POLITECNICO DI TORINO

Master's Degree in Computer Engineering Graphics and Multimedia

Master's Degree Thesis

Exploring the Potential of Generative AI in Media Production through Digital Humans

Supervisors

Prof. Andrea BOTTINO

Prof. Francesco STRADA

Ing. Roberto IACOVIELLO

Candidate

Valeria VALENTINI

December 2023

Summary

In recent years, *Synthetic Humans* and their applications have attracted considerable attention in a variety of fields, leading to an extensive exploration of their integration into Digital Twins, the Metaverse, and digital media production.

This thesis explores the complexities involved in the digital human production using a semi-automated approach to find a fair trade-off between high-quality outputs and efficient production times, which is critical in a small and agile context. The study was conducted in collaboration with RAI, using their photo and video archives to retrieve images of relevant subjects for texturing and 3D reconstruction.

The goal is to give *a second life to RAI's extensive archive* of 2D footage and propose improvements to the media experience. After an overview of the state of the art of Synthetic Humans creation, this study proposes innovative strategies in order to (i) automate the identified workflow involving *Unreal Engine 5* and *MetaHuman Creator*, and (ii) make it more versatile and modular.

In this work, the improvements have been distributed among different stages of the digital human creation process, starting with the scripted generation of 3D head meshes from 2D input images of the reference subject using a Blender plugin and then moving on to the generation of suitable images for texture development using *Stable Diffusion*, conditioned on the fine-tuning of the trained models. These assets are in turn integrated into the Unreal Engine, where a developed widget facilitates the posing, rendering, and texturing of the MetaHumans.
To complete the analysis, a thorough quantitative comparison between the subjects' original images and the rendered MetaHumans material to ensure an objective assessment of their similarity. In addition, subjective tests were performed to validate the chosen objective metric.

This work not only contributes to the field of Synthetic Humans and their application in the broadcast industry but also demonstrates the transformative potential of Generative AI in optimizing and enriching their creation workflow.

The insights and methodologies presented in this work lay the groundwork for significant advancements in creating realistic, versatile, and fully personalized Virtual Presences.

“Cursed creator! Why did you form a monster so hideous that even you turned from me in disgust? God in pity made man beautiful and alluring, after his own image; but my form is a filthy type of your’s, more horrid from its very resemblance”

Frankenstein, Mary Shelly

Table of Contents

Chapter 1

Introduction

Throughout history, the concept of artificial human beings has been a constant in the human imagination, emerging as a recurring theme that has fascinated different cultures and eras. From ancient mythological stories, such as the legend of Pygmalion, through the groundbreaking pages of Frankenstein written by Mary Shelley, the depiction of artificial humans has spanned time, coming to permeate modern science fiction and products of the entertainment industry.

These narratives and representations have played a crucial role, continually shaping and reformulating the way we perceive and imagine artificial humanity. Nowadays, the interest in this topic not only persists, but with the advent of Synthetic Humans, or Digital Humans, has intensified and the ambition to create human beings artificial humans are intertwined with the opportunities offered by the contemporary technological landscape.

The goal being pursued is dual-faced: firstly, to accurately replicate what is perceived as a "human being", and secondly, to transfer and reinterpret this essence within digital domains.

1.1 Understanding Synthetic Humans

Synthetic Humans, also known as Digital Twins of Humans, Embodied Conversational Agents, Virtual Agents or simply Avatars, are essentially digital representations of real individuals. These are not limited to merely reproducing physical appearance, facial expressions and body movements of a person, but, depending on their purpose, they can also extend to the analysis and reproduction of human's peculiarities that make each individual unique, such as personality, sensitivity, thoughts and abilities. They can recreate in digital space social components such as human behavior and communication, allowing empathetic interactions similar

to human ones.
Synthetic entities arise from the convergence of several disciplines, including computer graphics, computer vision, and artificial intelligence. Together with the use of Generative AI, it is possible to train them using specific information and data, generating a style or a "tone of voice" that reflects the user's needs, ensuring in this way a personalized and relevant communication.
In addition, thanks to sophisticated emotion-recognition algorithms, digital humans are emotionally receptive, able to perceive and respond to users' emotions, creating a bidirectional communication and demonstrating a level of empathy comparable to the human one.
The term "Synthetic Humans" emphasizes the aspiration to develop simulations so refined and detailed that they blur the boundaries between virtual and real, making these entities digital virtually indistinguishable from human beings.

1.1.1 Contextualization of the concept of Human Digital Twins

In recent years, the digital representation of humans, is subject to progressive interest, thanks to a decade-long alternation of the focus of digital twin computing between "things" and "human begins" that can be observed in figure 1.1.

Analyzing current trends, it is possible to predict that the digitization of humankind will shape the future, considering the Covid-19 pandemic that has accelerated the process of virtualization of services, the wide interest in VR/AR/MR experiences, and the commitment of Big Tech to the Decentralized Web3, in which the Metaverse and the consequent definition of a virtual identity will play a key role.
It must be acknowledged that the impact of digital humans will not be limited to the virtual world, as it is foreseeable that synthetic entities will surpass human capabilities in a number of areas, and the data generated by them will play a prominent role, increasingly replacing real data. Leading companies such as Microsoft, DataGen, Epic Games and Reallusion, are already adopting data from digital humans to enhance AI model training.[2] The goal is to overcome limitations of scarce, costly to collect, and privacy-sensitive real-world data, by enabling AI systems to learn from diverse, controlled and comprehensive synthetic datasets that mirror real human behaviors and interactions.

In the near future interaction with synthetic entities is set to become the norm in the emerging digital age, an evolution that will profoundly change the online experience and revolutionize technological, scientific and social fields.

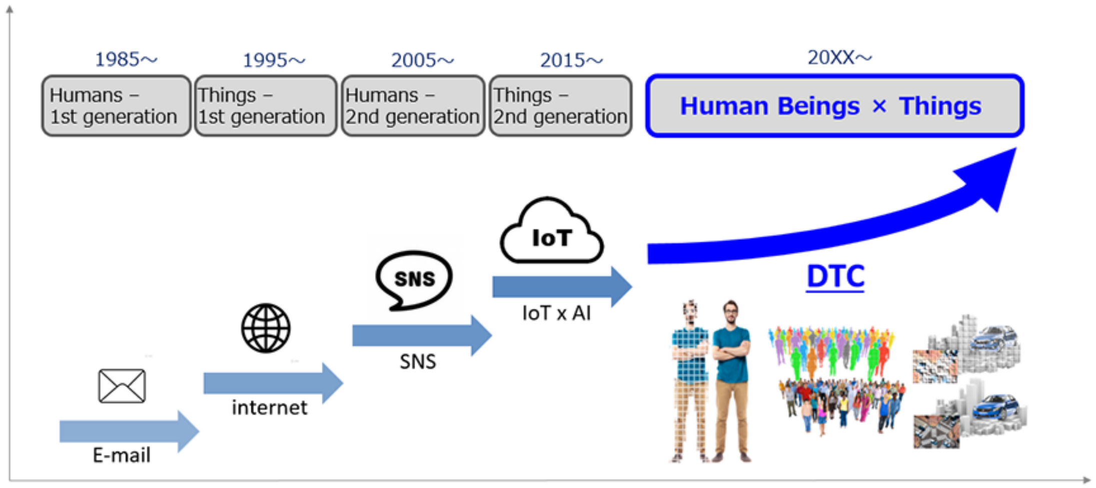


Figure 1.1: In 1985, human communication began to digitize with the introduction of e-mail. A decade later, in 1995, the focus shifted to the digitization of "things", such as time-schedules and maps, through the expansion of the Internet. In 2005, the focus shifted again to humanity, emphasizing connections and social networks through the emergence of social media. In 2015, digitization resumed its focus on objects, with the integration of the Internet of Things (IoT) and the development of artificial intelligence.[1]

1.1.2 Applications of Synthetic Humans in various sectors

Synthetic Humans find application in many domains, demonstrating their versatility and their cross-cutting impact:

- **Entertainment:**
  In the entertainment sector, video games stand out as a particularly rich domain for the application of these technologies, demonstrated by prominent titles like "The Last of Us" [3] and "Detroit Become Human" [4], in which digital characters express surprisingly detailed emotions and likeness, contributing to the creation of engaging and lifelike gaming worlds.
  Virtual production represents another area in which Synthetic Humans are valuable by simplifying and optimizing the creative process. The TV series "The Mandalorian" is an example of how the pre-visualization of scenes using digital actors can improve the efficiency of production. [5]
  The music and live performance world has also been transformed by the Synthetic Humans, with groundbreaking initiatives such as ABBA's virtual concert, which allowed fans to experience a unique live performance, even though in the physical absence of the band members, see figure 1.2.

In the social media field, digital personalities such as Lil Miquela [6] are virtual influencers and brand virtual ambassadors, have opened a new era in brand-consumer communication. Their ability to interact with audiences and participate in advertising campaigns marks a significant step toward new forms of engagement and digital relationships. Synthetic Humans are valuable tools for training in multiple fields such as aviation, medicine, education and many others. Their ability to reproduce realistic scenarios through digital human avatars results in more effective preparation of professionals and students, enabling them to deal with complex situations with greater confidence and safety.

Figure 1.2: Avatars on stage during the Abba Voyage concert in London.[7]

They take on the role of digital mentors, guiding new employees through the onboarding and on-the-job training process. This innovative approach accelerates the process of acquiring the required skills, while providing ongoing, personalized support.

In the educational field, they can take on the role of virtual tutors and adapt learning to the specific needs of each student, this includes simulating conversations in a foreign language, creating virtual labs, or reproducing realistic scenarios to facilitate the student immersion and enhance the learning experience.

Using Synthetic Humans to simulate the behavior and interactions of the workers within the Digital Twin environment, companies are able to assess risks, predict potential future scenarios, and identify opportunities for optimization. This approach not only allows companies to improve safety and operational efficiency, but also to refine the design of physical spaces in a precise way. Specifically, in architecture, their presence in Digital Twin models makes it possible to simulate the impact of design choices on the usability of spaces and the well-being of the people who occupy them.[8] This aspect turns out to be crucial for designing more functional, comfortable and safe buildings.

- **Scientific Research:** In the realm of scientific research, digital avatars facilitate the study of human behavior in controlled environments. Human perception, reactions and social interactions can be analyzed in detail, providing valuable information for a range of disciplines.
  In the fields of psychology and neuroscience, Synthetic Humans prove to be powerful tools for simulating human behaviors and studying emotional responses in specific situations. The creation of a "Virtual Identity" within virtual worlds opens up new perspectives for analyzing how people choose to represent themselves in digital contexts, offering valuable insights into the dynamics between virtual identity and self-perception.
  This aspect is particularly relevant in therapeutic settings, as it allows individuals to explore and confront aspects of their identity, fears or desires in a protected and controlled context, facilitating processes of introspection and self-knowledge

- **Health Care:** Synthetic Humans don't simply impersonate patients: they can also take on the role of doctors or medical staff, providing a mode of interaction that promotes empathy and the ability to relate to real patients.
  In terms of direct care, they can offer support during the post-operative recovery, facilitating both physical and cognitive rehabilitation processes.
  They are also able to perform an initial assessment of symptoms, that can direct patients to the most appropriate channels of care, the result is a more efficient patient management, that can avoid congestion in health facilities.
  Another relevant aspect is the creation of digital patient twins, which are faithful virtual replicas that can be used to monitor health status, predict the evolution of diseases and tailor treatments to the specific needs of each individual.

- **Communication and Interaction:** With the use of realistic avatars, digital communication gains depth and authenticity, elevating the quality of video conferencing, virtual chats and online interactions to a level never reached before.
  Through their ability to understand language and context, to faithfully represent a brand's identity and values, and to create authentic and empathetic interactions, Synthetic Humans offer brands new and effective ways of contacting and interacting with their audiences.
  This represents a significant advancement compared to traditional chatbots, such as those employed by Amazon, since unlike these conventional systems, Synthetic Humans are able to give context-aware responses that go far beyond simply answering a pre-set pool of questions.

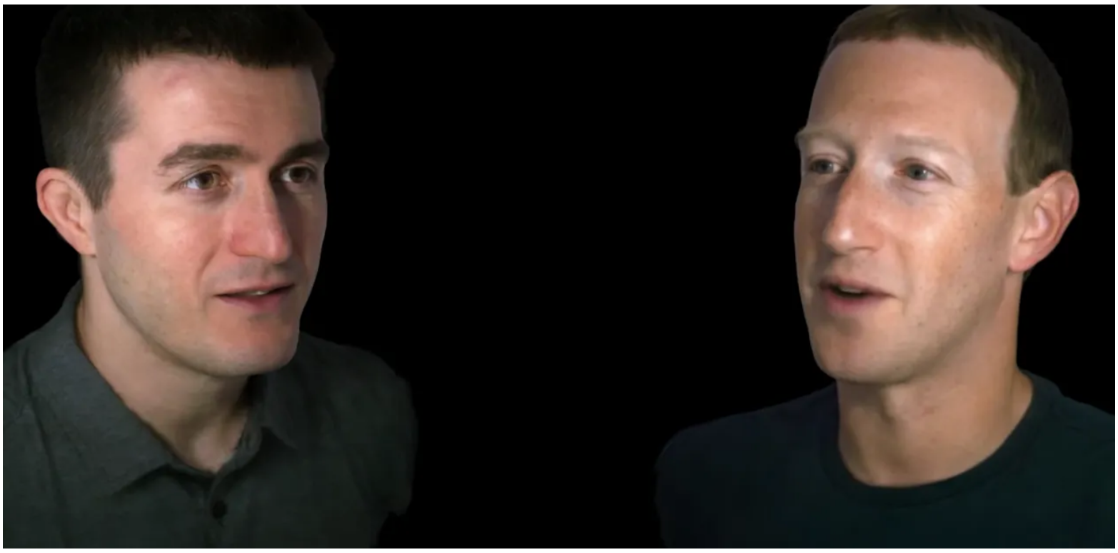


Figure 1.3: Mark Zuckerberg interviewed in VR using Meta's new Avatar Codecs[9]

1.2 Contextualization of the concept of Synthetic Humans in the broadcast world

In the broadcast industry, which embraces the transmission of audiovisual content through a variety of channels such as television, radio and online platforms, the inclusion of highly realistic digital characters is redefining how content is created, distributed and perceived, offering new levels of audience engagement and participation.

As television broadcasters and journalistic platforms [10] experiment with the use of digital avatars to conduct specific programs and segments, **virtual interviews** can be conducted, in which a real person converses with a synthetic character. This mode proves particularly valuable for interviewing people located in geographically distant places and times. In parallel in the context of **livestreaming and real-time conversations** [11], Synthetic Humans open new frontiers of interaction.

Whether at sporting [12] events or live events, a "Real Time Conversation Digital Human" can take on the role of conductor, providing live commentary and interacting with the audience in a dynamic and engaging manner.

In the field of **documentary and film production**, the use of Synthetic Humans results in the possibility of exploring new visual narratives. The ability to recreate historical characters opens up new storytelling perspectives, allowing content creators to bring moments and figures from the past to life, while their presence in films and TV series as special effects allows them to overcome physical and creative limitations, replacing human actors in scenes of particular complexity.

They can be used to create virtual avatars for **advertising campaigns**, allowing greater flexibility and creative control. These avatars can be programmed to interact with consumers in personalized ways, enhancing the shopping experience and providing relevant product information. [13]

1.2.1 Impact of Synthetic Humans on the Future

The future of broadcasting promises to be intrinsically connected to Synthetic Humans, with intriguing prospects for entertainment. Viewers will have the opportunity for more immersive television and film experiences opening up new dimensions of engagement and participation.
Entertainment will become more interactive and personalized, adapting to viewers' individual preferences. In parallel, accessibility in media will reach new heights, they will offer innovative solutions for people with disabilities, providing more comprehensive access to television programs, multimedia content and information. Avatars will be able to communicate in sign language, support visual accessibility, and offer new ways of enjoying content for people with developmental disabilities. This will represent a significant step toward a more inclusive society.

In addition, Synthetic Humans will serve as a bridge between physical reality and the Metaverse, becoming an essential part of the interaction between the physical environment and the virtual world. The Metaverse could become a platform where viewers interact directly with Synthetic Humans in virtual shows and experience digital worlds.

1.3 3D Reconstruction of Human Models: A Research Perspective

In recent years, significant studies have been conducted in the realm of three-dimensional reconstruction of human models. The three-dimensional reconstruction of digital models of humans poses a complex challenge within the fields of computer vision and graphics. The primary objective is to accurately retrieve the geometry and appearance of humans from visual sources such as images, videos, or depth data, precisely translating two-dimensional information into detailed and realistic three-dimensional representations of human bodies.

This section of the chapter explores the latest research [14] in this domain, highlighting innovative methodologies and techniques that contribute to the refinement of creating increasingly realistic and detailed avatars.

1.3.1 Detailed and Realistic Representation

The primary focus is on creating three-dimensional models capable of accurately capturing not only the physical form of an individual but also subtle details such as facial expressions and clothing folds. Challenges arise from anatomical complexity and the need to make such models accessible, efficient, and cost-effective.

1.3.2 Evolution from Traditional Pipelines to Machine Learning Innovation

Traditional reconstruction pipelines, relying on complex acquisition systems, serve as a fundamental starting point. However, limitations in terms of cost and portability of such approaches have led to the adoption of innovative machine learning-based approaches. These new approaches aim to overcome intrinsic challenges by providing efficient, accessible solutions capable of producing photorealistic three-dimensional models.
Different approaches to reconstructing the 3D model are illustrated in Figure 1.4:
a) The traditional reconstruction pipeline necessitates a dense camera array or depth cameras and involves numerous isolated steps.
b) Regression-based methods employ neural networks to directly regress human geometry or appearance from input images.
c) Optimization-based methods, utilizing differentiable rendering, reconstruct 3D human models by minimizing the rendering error between re-rendered images and the corresponding input images.

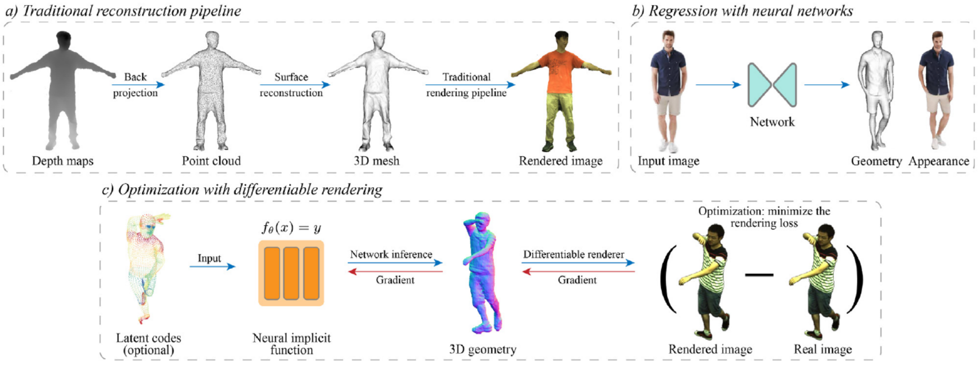


Figure 1.4: Comparison between three reconstruction approaches [14].

1.3.3 Current Trends and Developments

Recent advancements indicate a shift towards using deep neural networks to enhance the efficiency and robustness of reconstruction by learning human models from existing data. Implicit function-based approaches emerge as a preference over traditional forms such as meshes and voxels, emphasizing the importance of detailed visual rendering [14].

Mesh-based Approaches

Mesh-based approaches play a crucial role in three-dimensional human reconstruction, enabling the creation of detailed models that accurately reflect facial expressions and clothing shapes. However, the low dimensionality of meshes may limit their ability to capture high-frequency details [14].

Implicit Function-based Approaches

The use of implicit functions proves to be an effective alternative, with advantages such as ease of optimization, spatial resolution independence, and memory efficiency. Techniques like "Pixel-aligned Implicit Function (PIFu)," "PIFuHD," and "Geo-PIFu" enhance the rendering of details and local features [14].

Differentiable Rendering

Differentiable rendering is crucial for achieving highly photorealistic human models. Challenges and techniques, such as "mesh renderer," "differentiable sphere tracing," and "volume rendering," are explored to address efficiency issues and dynamic scene modeling [14].

1.3.4 Future Perspectives

Future research focuses on generalizable methods, efficient reconstruction for multiple subjects, advancements in photorealistic rendering, and improved techniques for scene content manipulation, such as object manipulation and clothing changes. The innovative use of neural networks promises to make three-dimensional reconstruction more accessible and cost-effective, thereby overcoming limitations of traditional hardware-intensive setups.

1.4 Problems and Challenges

This innovation of Synthetic Humans brings significant challenges and technical problems that require strategic solutions.

Exploring key issues and challenges associated with the utilization of Synthetic Humans as digital representations in the broadcast world will be undertaken.

- **Data Integration and Quality** Creating an accurate photorealistic avatar requires the integration of data from various sources, including motion suits, mobile cameras, different software, and existing archives. Ensuring the quality, consistency, and reliability of this data can be a significant challenge. Data can be noisy, incomplete, or discordant, making it essential to develop rigorous data acquisition protocols and quality control systems.

- **Interoperability** Different workflow components may use different technologies and standards, making it complex to create an integrated pipeline for building Digital Humans.
  Ensuring that all components communicate effectively and seamlessly can be a daunting task. Standardization of interfaces and protocols can help mitigate this challenge.

- **Model Validation** Developing accurate 3D models that closely mimic human behavior is essential for the creation of convincing Digital Twins.
  Validating these models to match actual performance in terms of similarity and movements can be difficult due to the lack of universal metrics or standards for assessing realism.

- **Cost and Resource Allocation** Developing a high-fidelity Synthetic Human can be resource intensive, requiring time, money, and people involved in the project.
  Adopting AI-based software and tools can be a strategic challenge, as it requires planning and allocating resources effectively to achieve the desired results.

- **Lack of Standardization** The lack of standardized protocols for the development of Digital Twins can lead to compatibility issues and difficulties in collaboration across different organizations or sectors.
  Establishing common standards and shared protocols could help overcome these barriers.

- **Cultural and Organizational Resistance** The introduction of new technologies for creating photorealistic avatars may require changes in the way production centers work and make decisions. There may be cultural or organizational resistance from employees who are unfamiliar with the technology or reluctant to adopt new processes.

- **Business Aspects** The lack of standardized protocols for rewarding the work of talent and ensuring the quality of work performed can create obstacles and risks. Establishing clear rules and commercial guidelines could be crucial to sustaining the Synthetic Humans industry.

- **Ethics and Privacy** Regarding the creation and manipulation of digital representations of real individuals, strict principles need to be established to ensure that the dignity and rights of digitally represented individuals are respected. Distributed content using digital avatars must be clearly distinguishable and identifiable from real ones, the goal is to maintain transparency so as to safeguard both the integrity of the individual and to build a more ethical and responsible digital environment.

1.5 Purpose and Objectives of the Thesis

This thesis work, conducted in collaboration with RAI, the well-known Italian radio and television broadcaster, aims to analyze and improve the workflow for the production of photorealistic avatars that portray relevant figures, such as historical figures, performers, and athletes, by utilizing the resources available in the RAI archive.

The initial phase of the research focused on outlining every stage of the workflow, with the aim of establishing a comprehensive and organized set of procedures for developing Synthetic Humans. By conducting a thorough examination of each phase, the main technical and creative challenges that emerge during the avatar generation process were identified.

Afterwards, an evaluation of the generated Synthetic Humans was conducted, with an emphasis on fidelity and likeness to the original subjects. The goal was to obtain the most accurate outcomes while minimizing the resources expended in terms of time, people and artistic skills required.

In conclusion, the thesis is dedicated to identifying workflow automation solutions to make the entire process more efficient and reduce the need for human involvement. This work is proposed as a first step forward in the field of automatic generation of photorealistic avatars, with the aspiration of making the integration of Synthetic Humans in television and multimedia productions more accessible and flexible.

Chapter 2

State of Art

To embark on the path towards the automated creation of photorealistic avatars from historical images in the RAI archive, it is essential to gain a comprehensive understanding of the state of the art in three critical areas:

- Image restoration
- 3D facial reconstruction
- Facial animation

This chapter is dedicated to examining advanced software solutions and prevailing methodologies in these three domains, with the aim of providing a comprehensive overview of current trends and best practices.

Furthermore, a special emphasis will be placed on MetaHuman, a cutting-edge technology in the realm of digital character creation, given its increasing adoption in the entertainment and media industry.

2.1 Image Restoration

The image restoration process aims to obtain high-quality images from damaged or compromised input images. Image corruption can occur due to the capture process (e.g., noise, lens blur), post-processing (e.g., JPEG compression), or photography under non-ideal conditions (e.g., fog, motion blur).

Historical images from the RAI archive often suffer from issues such as low resolution and noise, making the use of Image Restoration tools essential.

2.1.1 Adobe Photoshop Super-Resolution feature

Adobe Photoshop Super-Resolution [15] is a well-known software for upscaling and Super Resolution of images using machine learning. Through machine learning, it analyzes image patterns and generates realistic and consistent details during upscaling.

It is the most widely used and versatile software because it offers a high degree of creative control, advanced retouching tools, and the option to restore textures manually to ensure optimal results.
Its processing speed allows for a seamless workflow.

2.1.2 Topaz Gigapixel

Topaz Gigapixel [16] is specialized software for upscaling images and offers detailed control over resolution and quality. It is designed to increase the resolution up to 6 times the original image size. Unlike Adobe Photoshop, it allows you to preview the image before upscaling.
Because it is specialized for this task, it provides more controls for addressing issues like blurring, noise, color bleed, and face refinement.

2.1.3 GFP-GAN

This project is built upon a specially trained Generative Adversarial Network (GAN) for facial image restoration, with an emphasis on prioritizing facial details. GFP-GAN [17] distinguishes itself for its ability to achieve an optimal balance between realism and fidelity in the image restoration process. This achievement is made possible through the utilization of a broad and diverse knowledge base embedded in a pre-trained generative network, specifically designed for facial image restoration.

As we can see in figure 2.1, it consists of a degradation removal module (U-Net) and a pre-trained face GAN serving as a facial prior [17]. These components are connected through latent code mapping and various levels of Channel-Split Spatial Feature Transform (CS-SFT). During training, the following processes are used:
1) Intermediate restoration losses are employed to remove complex degradation.
2) Facial component losses with discriminators enhance the face.
3) Losses for preserving identity are applied to maintain the identity of the face.

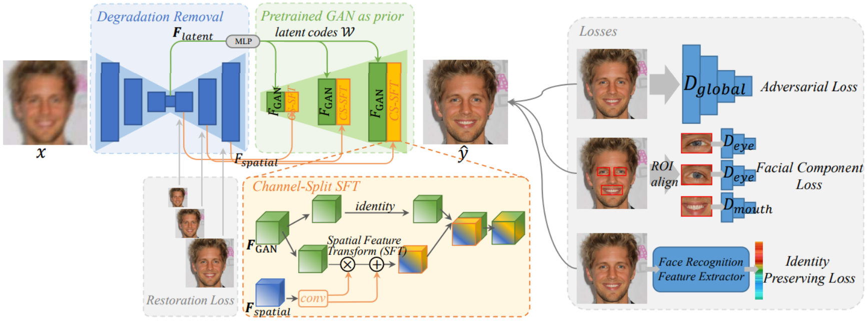


Figure 2.1: The GFP-GAN framework overview

2.1.4 CodeFormer

In the quest for solutions to restore damaged or dated faces, CodeFormer [18] emerges as a powerful facial restoration algorithm.
CodeFormer reduces uncertainty in transforming degraded faces into high-quality ones. This method is based on a Transformer-based prediction network to model low-quality faces and predict codes, producing impressive results even with heavily damaged images. Thanks to its advanced knowledge and global model, CodeFormer outperforms existing methods in terms of quality and fidelity, demonstrating significant robustness to degradation. Results on both synthetic and real-world datasets confirm the effectiveness of this approach.

As we can see in figure 2.2 initially, a discrete codebook and a decoder are learned to store high-quality visual parts of facial images through self-constructive learning. Then, with a fixed codebook and decoder, a Transformer module is introduced for predicting the code sequence, modeling the global composition of the low-quality input face. Additionally, a controllable feature transformation module is used to control the information flow from the low-quality encoder to the decoder.
Note that this connection is optional and can be disabled to avoid adverse effects when inputs are severely degraded, and a scalar weight can be adjusted to strike a balance between quality and fidelity. [18]

2.2 Face Reconstruction

3D Face Reconstruction, which also includes **Texture Reconstruction**, is a computer vision activity involving the creation of a 3D model of a human face from a 2D image or a set of images. The output of this process can be used for various applications such as virtual reality, animation, and biometric identification.

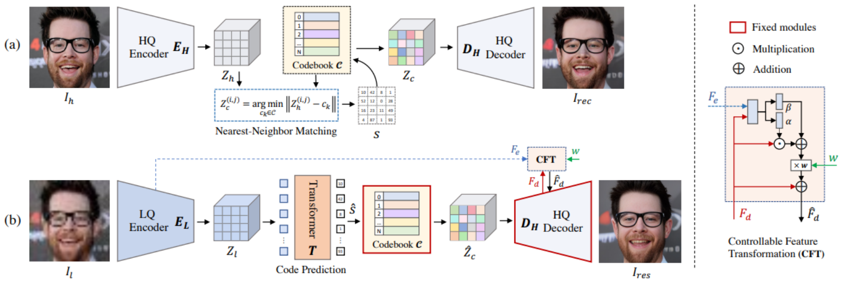


Figure 2.2: CodeFormer Structure

This section will explore both advanced methodologies and software/tools/plugins for creating 3D models of human faces from 2D images or sets of images.

2.2.1 3D Photogrammetry

Among the most commonly used methodologies for face reconstruction, we have **3D Photogrammetry**.
This approach involves capturing 2D images from various angles and using stereo vision algorithms to generate 3D models of faces. Examples of software based on this technology include ***Intel RealSense*** [19] and ***RealityCapture*** [20].

Intel RealSenseis an Intel 3D sensor platform that captures high-quality 3D data from human faces using RGB cameras, infrared sensors, and a structured light projector. The associated software processes this data in real-time to create 3D models of faces that can be used in facial recognition and animation applications.

RealityCapture is photogrammetry software that generates 3D models, including facial models, from 2D images captured from various angles. This software is known for producing high-resolution 3D models with realistic textures and is widely used in industries such as architecture, video games, and film.

2.2.2 3D Scanning

The 3D scanning technique for Face Reconstruction is an advanced method for acquiring detailed three-dimensional data of the human face. This approach involves using 3D scanning devices, such as Microsoft Kinect or similar scanners, that employ structured light technology, which utilizes the deformation of projected light on a face to calculate its 3D geometry.

This technology is widely employed in facial modeling, offering a precise solution for capturing facial geometry. A well-known software that works with the Microsoft Kinect is ***Skanect*** [21], which processes 3D scanning data to create detailed facial models usable in applications such as animation, virtual reality, and game development.

2.2.3 Blender Face Builder plugin

Face Builder is a plugin developed by Keentools for Blender, and it provides an intuitive tool for creating 3D models of human faces using a series of reference photos. [22] [23]

FaceBuilder reads the information contained in the photos and uses it to automatically set up virtual cameras, allowing you to use photos of various sizes, like in figure 2.3.

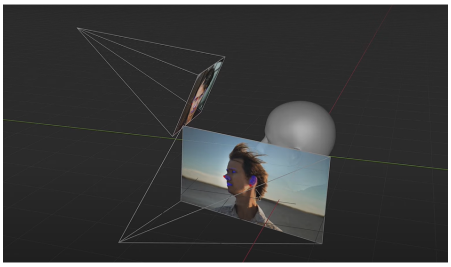


Figure 2.3: Automatic camera setup, detection, and estimation of the format.

For the highest quality and precision, it is recommended to use photos with neutral facial expressions, as we can see in figure 2.4. However, FaceBuilder is also capable of supporting non-neutral facial expressions, allowing for acceptable results even with photos that do not show neutral expressions.

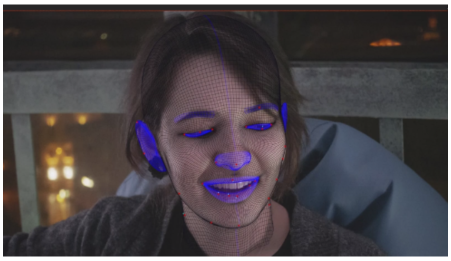


Figure 2.4: Support for facial expressions.

The use of artificial intelligence automates the process of aligning facial points, like in figure 2.5, eliminating the need for manual point placement initially.
However, it's important to note that automatic alignment may not always be perfect, so slight manual adjustments may be required. Nonetheless, this approach significantly reduces the time needed for facial modeling.

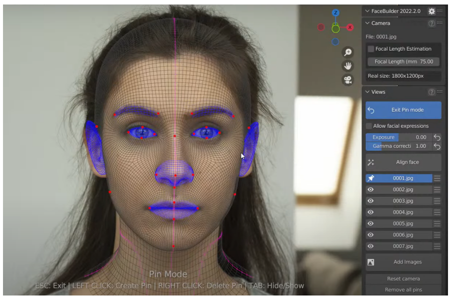

File: 0001.jpg
Focal Length (mm 75.00
Real size: 1800x1200px
Exposure 0.00
Gamma correcti 1.00
ESC: Exit | LEFT CLICK: Create Pin | RIGHT CLICK: Delete Pin | TAB: Hide/Show

Figure 2.5: Automatic face alignment with AI.

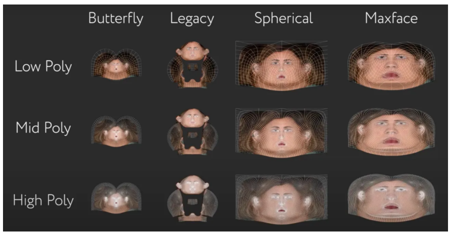


Figure 2.6: UV maps in Face Builder. [23]

It allows for extracting the texture for the 3D face using the views in which the model has been alligned. FaceBuilder offers four UV map options, that we can see in figure 2.6, and there are: Butterfly, Legacy, Maxface, and Spherical, each with specific features. In the latest version, a UV map (mh that stands for MetaHuman) compatible with Metahuman has been made available.

Models generated through FaceBuilder can be easily exported in formats compatible with Unreal Engine and Unity, enabling their use in game development and animation contexts. This gives FaceBuilder a high degree of versatility in terms of functionality and export capabilities, making it an extremely valuable resource.

Another advantage of FaceBuilder is its ability to import facial animations recorded using the Live Link Face app by Epic Games. This app is compatible with creating automated animations on Metahuman through Unreal Engine.

2.2.4 Character Creator Headshot plugin

The Headshot plugin for Character Creator [24] serves as an alternative to Blender's Face Builder. It provides advanced features for generating realistic 3D models from photographs, with the ability to directly manipulate the face shape and a wide range of customization options through control areas directly on the 3D model.

Furthermore, the model generated by Headshot already includes a complete rig for voice lipsync animations, facial expressions, and body animations.

The term "complete rig" refers to a set of virtual bones that can be controlled, allowing realistic animation of the 3D model. The complete rig for voice lipsync animations is designed to synchronize the movements of the model's lips with the audio, while facial expressions can be modified and customized using the controls included in the rig.

Additionally, this plugin offers solid compatibility with software such as Unreal, Blender, and NVIDIA Omniverse.

2.2.5 Daz3D Face Transfer feature

This is a feature of Daz Studio, a 3D modeling and rendering software specialized in character creation [25].

This feature allows you to use a single photo for a face scan, and after a calibration process of key points, these will be transferred to a basic 3D model. After creating the 3D face model, you can further customize it by adjusting parameters such as face shape, distinctive features, and expressions. However, it requires high-quality images for optimal results.

Additionally, it is compatible with the NVIDIA Omniverse pipeline.

2.2.6 Avatar SDK

Avatar SDK [26] utilizes artificial intelligence (AI) for the creation of 3D avatars from images. It provides advanced customization options, including face and body shape, and allows for adjustments of facial features to achieve realistic results.

2.2.7 FaceGen

FaceGen [27] is a software for creating realistic 3D models of faces and heads. It offers numerous controls for modifying facial features and detailed rendering.

2.2.8 AVATAR ME++

AVATAR ME++ [28] represents an advanced method for the reconstruction of photorealistic 3D faces from individual images. This project introduces a rendering methodology that allows for realistic output even in different lighting environments. It uses Albedo and Normal maps to make the reconstructed faces highly detailed, capturing the facial skin's appearance, shadows, and reflections accurately. This method promises to capture precise facial appearance details, making 3D avatars extremely realistic and suitable for use in virtual environments and animation applications.

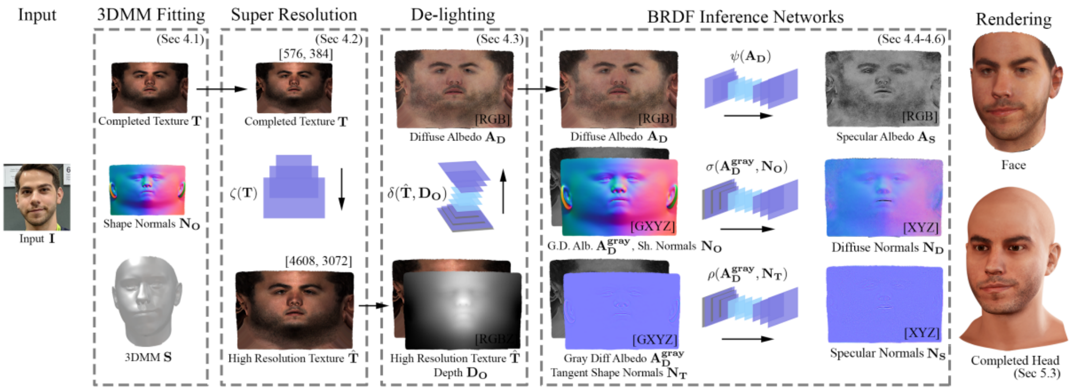

Input
3DMM Fitting
Super Resolution
De-lighting
BRDF Inference Networks
Rendering
(Sec 4.1)
(Sec 4.2)
(Sec 4.3)
(Sec 4.4-4.6)
Input $\mathbf{I}$
Completed Texture $\mathbf{T}$
Shape Normals $\mathbf{N_O}$
3DMM $\mathbf{S}$
[576, 384]
Completed Texture $\mathbf{T}$
$\zeta(\mathbf{T})$
[4608, 3072]
High Resolution Texture $\hat{\mathbf{T}}$
[RGB]
Diffuse Albedo $\mathbf{A_D}$
$\delta(\hat{\mathbf{T}}, \mathbf{D_O})$
[RGBZ]
High Resolution Texture $\hat{\mathbf{T}}$
Depth $\mathbf{D_O}$
[RGB]
Diffuse Albedo $\mathbf{A_D}$
[GXYZ]
G.D. Alb. $\mathbf{A_D^{gray}}$, Sh. Normals $\mathbf{N_O}$
[GXYZ]
Gray Diff Albedo $\mathbf{A_D^{gray}}$
Tangent Shape Normals $\mathbf{N_T}$
$\psi(\mathbf{A_D})$
$\sigma(\mathbf{A_D^{gray}}, \mathbf{N_O})$
$\rho(\mathbf{A_D^{gray}}, \mathbf{N_T})$
[RGB]
Specular Albedo $\mathbf{A_S}$
[XYZ]
Diffuse Normals $\mathbf{N_D}$
[XYZ]
Specular Normals $\mathbf{N_S}$
Face
Completed Head
(Sec 5.3)

Figure 2.7: AVATAR ME++ Architecture is a 3D Morphable Model (3DMM) adapted to an "in-the-wild" image to create a detailed UV texture. Neural networks process normals, albedo, and face shape to realistically render the face and head in various situations. [28]

2.2.9 DAD-3D Heads

The DAD-3D Heads project [29] represents a valuable resource for 3D facial reconstruction. It is a dense, accurate, and diverse dataset designed for three-dimensional head alignment from a single image. This dataset contains over 3,500 landmarks that accurately represent the three-dimensional shape of the head. The proposed model uses this dataset to learn head shape, expression, and pose parameters. It then employs a Fits Like A Glove (FLAME) mesh for three-dimensional reconstruction. This approach offers a high degree of accuracy and diversity, making it a promising choice for 3D facial reconstruction, although it does not address texture reconstruction.

As we can see from the figure 2.8, the Gaussian estimator predicts the coarse positions of head landmarks. The fusion block combines the coarse heatmap, the BiFPN feature map, and the output of the CNN encoder to adjust a series of parameters of the 3D head model and the finer positions of head landmarks. [29]

2.2.10 OSTEC

OSTEC [30] is a project that focuses on facial texture reconstruction without the need for extensive facial texture datasets. Instead, it leverages the knowledge

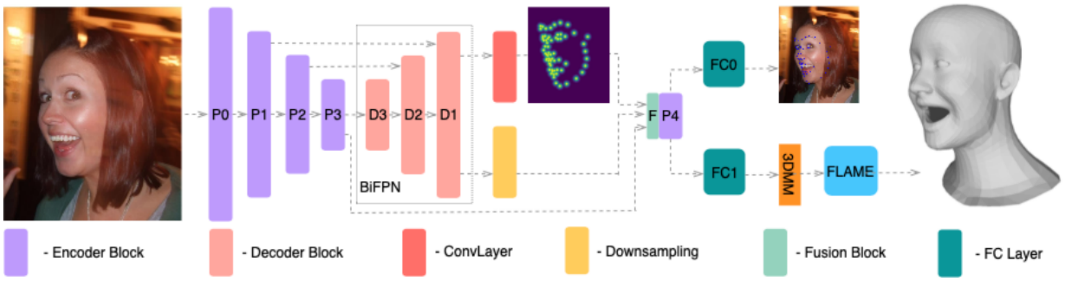


Figure 2.8: DAD-3DNet Architecture Design

stored within the algorithm to achieve high-quality results. The process involves 3D rotation of the input image to obtain different views of the face. Subsequently, a 2D face generator is used to reconstruct the parts of the image that are visible in these different views. This 2D face generator relies on the knowledge stored in generative models to produce a realistic representation of the observable parts of the face. Experiments demonstrate that the completed UV textures and front-facing images are of high quality and maintain the original appearance of the facial identity. This approach offers an innovative solution to facial texture reconstruction. It is possible to observe its architecture in the figure 2.9.

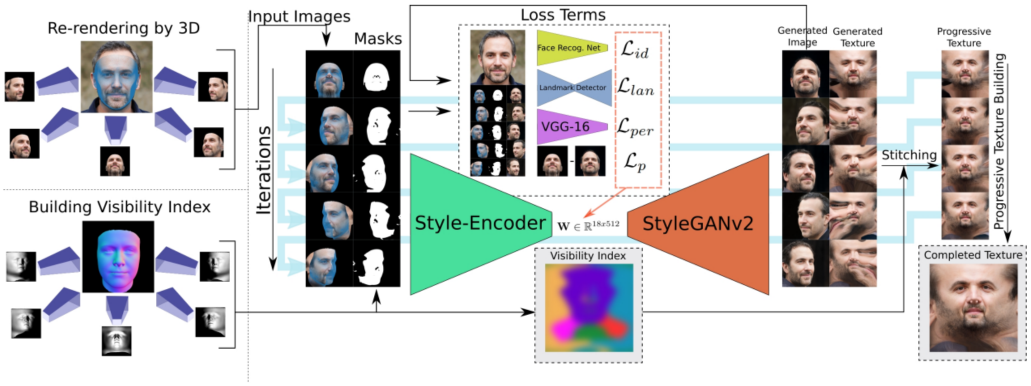


Figure 2.9: OSTEC Architecture The proposed approach iteratively optimizes the UV texture maps for various re-rendered images along with their masks. At the end of each optimization, the generated images are used to acquire partial UV images from dense landmarks. Finally, the completed UV images are sent to the next iteration for the progressive construction of textures. [30]

2.3 Facial Animation

Facial animation is crucial for bringing life and realism to photorealistic avatars. There are various facial animation techniques, including keyframe animation, motion capture-based facial animation, and animation based on artificial intelligence algorithms like neural networks.

These techniques involve the use of animation software. Below is a list of prominent software for facial animation.

2.3.1 iClone

iClone [31], developed by Reallusion, uses motion capture to create realistic facial animations. It can record the movement of an actor and apply it to a virtual character in real-time, ensuring accurate lip synchronization (lypsync animation) and emotional expressions.

2.3.2 Audio2Face

Developed by NVIDIA, it's an AI-based facial animation software that allows for automated facial animations generated from audio recordings [32].
The input audio is then fed into a pre-trained neural network, and the output will drive the 3D vertices of the character's mesh, creating real-time facial animation.

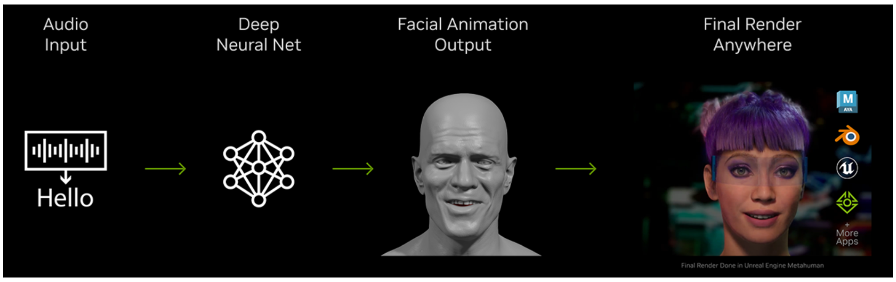


Figure 2.10: From audio to animation with ease, thanks to generative AI. [32]

2.3.3 Faceware

Faceware is a company that has developed markerless facial mocap software that works in real-time and allows for precise capture of an actor's facial expressions and transferring them to 3D models.
It is widely used in the film and video game industry to capture realistic facial performances. [33]

2.4 Creation Avatar Softwares

In the context of creating photorealistic avatars, the current landscape offers advanced platforms and software solutions, with **MetaHuman** and **Character Creator** being of the leading options.

2.4.1 Metahuman

MetaHuman is a technology developed by Epic Games, representing a significant step in the generation of exceptionally high-quality and realistic digital characters.

MetaHuman Creator [34], the core component of MetaHuman, is at the heart of this innovation. This platform, accessible through a cloud-based application, allows for the generation of photorealistic avatars in a matter of minutes. This is a remarkable breakthrough from traditional methods, enabling content creators to quickly and intuitively create highly realistic digital human characters.

The efficiency of MetaHuman Creator extends to its flexibility. Users can start with predefined character presets or import custom meshes using the Unreal Engine plugin. In fact, the production pipeline for a MetaHuman revolves around MetaHuman Creator, which allows for creating various characters by combining different predefined features.

These MetaHumans created in MetaHuman Creator can be imported into Unreal Engine projects via Quixel Bridge, a plugin directly integrated into UE5.
This flexibility provides endless creative possibilities, giving developers and artists the freedom to bring their own visions to life.

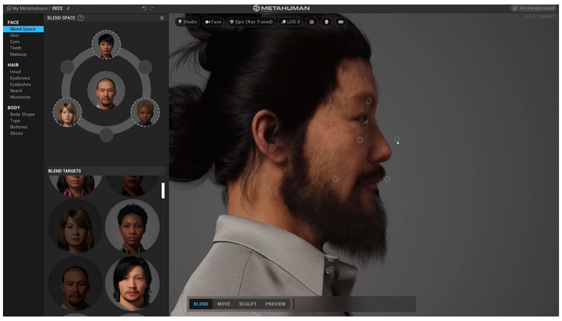


Figure 2.11: MetaHuman Creator

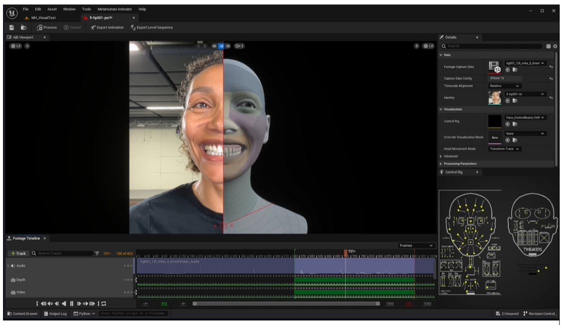


Figure 2.12: MetaHuman Animator

MetaHuman Animator [35] is another key component designed to capture high-fidelity facial performances on any MetaHuman character. This tool is capable of capturing every nuance of actor performances and convincingly transferring them to digital characters. This results in engaging and believable experiences for the audience.

Facial animation with MetaHuman Animator does not require complex equipment. Common devices like the iPhone can be used to achieve high-quality results, making the process accessible to a wide range of creators. This flexibility in performance capture allows for adapting the hardware to the specific needs of the project, significantly improving the process of creating MetaHumans and customized digital characters.

MetaHuman offers an efficient workflow that can be easily integrated into any existing content creation process. Creating a MetaHuman takes only a few minutes, significantly simplifying the digital character development process. Furthermore, MetaHumans can be exported to Digital Content Creation (DCC) software like Maya and Blender for further customization and made ready for real-time use in Unreal Engine, opening up new creative possibilities for content creators of all kinds.

MetaHuman goes beyond character creation, offering advanced capabilities for integrating realistic elements. Through the implementation of ray tracing and simulation models like subsurface scattering, it is possible to simulate details such as hair, fur, eyelashes, beard, and even achieve a high level of color fidelity, utilizing textures for wrinkles and blood flow beneath the skin.

2.4.2 Character Creator

Character Creator [36], developed by Reallusion, stands out as a versatile and comprehensive solution for shaping high-quality digital characters, aligning itself with the user-friendly approach of MetaHuman. This technology distinguishes itself through an intuitive interface that streamlines the creation of detailed and lifelike characters.

The flexibility of Character Creator is evident in its ability to adapt to user needs. Users can either kickstart their designs with predefined presets or import custom meshes seamlessly using dedicated plugins.

Furthemore, The platform excels in crafting photorealistic details, managing elements like hair, fur, eyelashes, and beard with advanced tools. The platform provides an authentic reproduction of textures for wrinkles and blood flow beneath the skin, contributing to the authenticity of characters.

Character Creator seamlessly integrates into the pipelines of other leading tools, allowing users to effortlessly import characters into projects like Unreal Engine 5 and Unity. Additionally, Character Creator facilitates integration with digital content creation (DCC) software such as Maya and Blender, enabling further customization and preparing characters for real-time use.

Designed to deliver engaging experiences, Character Creator enables the creation of characters capable of conveying highly realistic facial expressions. A notable aspect is the ability to animate the face through ***iClone***, another software by Reallusion that, without the need for complex equipment, allows for high-quality results.

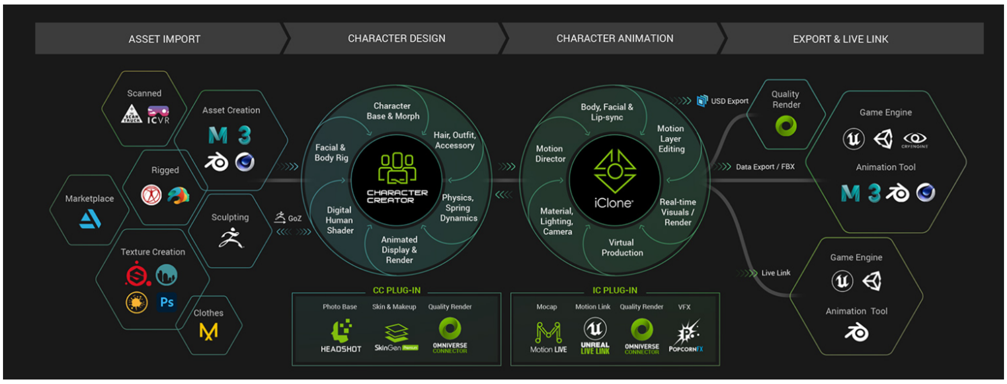


Figure 2.13: ***iClone*** and ***Character Creator*** seamlessly work together as one big design-to-animation character platform and greatly ease the in-and-out productivity with many other 3D applications.

In summary, Character Creator is a robust choice for crafting photorealistic avatars, seamlessly integrating into workflows with its rich features and flexibility. Its intuitive interface and exceptional detailing make it a standout player in digital character creation.

Chapter 3

Technologies Involved

In this thesis, the use of Generative AI is explored to simulate a realistic face scan environment with the aim of generating subject's rendering that are suitable for the face texture mapping.
By employing Stable Diffusion [37], a generative model, is possible to produce images where the subject is positioned facing forward and at 45-degree angles, recreating standard poses used in head photogrammetry capture.
This method also allows for precise control over the image generation process, enabling the exclusion of undesired elements like hair or glasses and the adjustment of attributes such as age and lighting conditions. The result is a set of idealized images, ready for texture extraction, eliminating the need for the physical presence of the subject.

3.1 Generative AI

Generative AI, or GenAI [38], refers to the use of artificial intelligence for the creation of new, synthetic data, which can include a variety of formats, like imagery, audio, code or natural language.
While discriminative models excel in making predictions or classifications based on input data, generative models take a step further by exploring the mechanisms behind the creation of entirely new data instances.
At the core it operates through Machine Learning, and learns patterns and relationships within large datasets of existing content, through this learning process, it develops a statistical model.
When given a prompt, GenAI employs the model to generate an output that represent what an expected response might be, this process leads to the production of new, original content that references the original data but is unique in its composition.

For image synthesis, there are several types of models which are trained on large datasets of images. The most commonly used generative models are:

- **Generative Adversarial Networks:** GANs [39] consist of two competing neural networks, a generator and a discriminator.
  The former creates images that closely resemble training data, aiming to produce convincing fakes, while the latter has the task of differentiating between the generated images and the actual training images, penalizing the generator for any implausible results. This dynamic creates a competitive environment where both networks continuously learn and improve based on mutual feedback, enhancing the overall quality of the generated images.
  GANs have found widespread use in enhancing image resolution, style transfer, creating deepfake videos, and generating images for applications like "This Person Does Not Exist". [40]
  However, they are also known for potential instability and less diversity in generation due to their adversarial training nature. A notable issue is Mode Collapse, [41] it happens when the generator produces a limited variety of outputs, often replicating the same image, because it identified a specific output that consistently deceives the discriminator. Consequently the generator focuses on producing only that output, and it leads to a limited diversity in generated results, reducing the GAN's ability to capture the full complexity of the intended data distribution.

- **Diffusion Models:** This models [42] in the recent past years have gained increasing interest [43] and are inspired by the principles of thermodynamics and the diffusion process, which involves the movement of particles from region of higher concentration to lower concentration. Their versatility is showcased through two distinct modes of image creation: Unconditioned and Conditioned generation [44].
  Unconditioned Generation operates without any external input. It's particularly effective for tasks like synthesizing human faces or achieving super-resolution in images, relying on learned data distributions.
  Conditioned Generation, incorporates external inputs such as text descriptions or partial images. This mode is widely used in text-to-image generation or image inpainting, allowing the model to fill in missing parts of an image or to create images that are coherent with textual descriptions.
  The core of the training procedure involves two phases, the **Forward Diffusion Process** and the **Reverse Diffusion Process**.

  An high-level overview of the Forward diffusion process begins with the original training image x_0 and over a series of steps, this image is gradually corrupted by adding Gaussian noise. This noise is not added all at once but rather

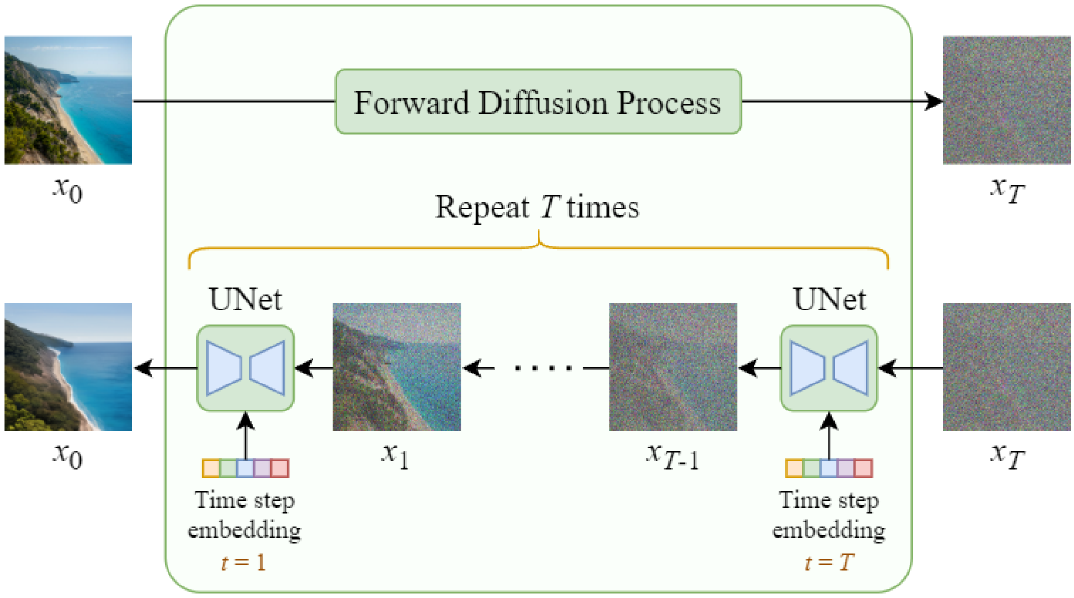


Figure 3.1: Pure diffusion model architecture [45]

incrementally over a series of steps T. Each step adds an additional layer of noise, resulting in a series of images that are increasingly noisy $(x_1, .., x_T)$. By the end of this process, the image is often indistinguishable from pure noise. The latter phase, known as the Reverse diffusion process, is characterized by undoing the Forward diffusion process and is executed with the support of a trained neural network UNet [46]. This network is adept at predicting the noise added to each image by employing time step embeddings, which provide contextual information about the order in which noise was added. This information guides the network to accurately and successively subtract the predicted noise from each image in the series.
From Figure 3.1 the process begins with the fully noised image x_T, the UNet model operates as a loop, and employs its predictions to sequentially remove the noise. At each step, it reverses the noise addition from the forward process, progressively getting closer to the original image.

3.1.1 Stable Diffusion

Released in August 2022, Stable Diffusion has achieved state of the art tool for image generation and suited for the aim of the project that requires high-quality image generation with specific control and adaptability requirements. Stable Diffusion is a Latent Diffusion Model (LDM), a novel approach proposed in the paper "High-Resolution Image Synthesis with Latent Diffusion Models" [47]. As shown in Figure 3.2 it involves the forward and backward processes used in Diffusion models, but it operates on a compressed representation of images, opposed to working directly in the pixel space with is computationally very slow. A Variational Autoencoder

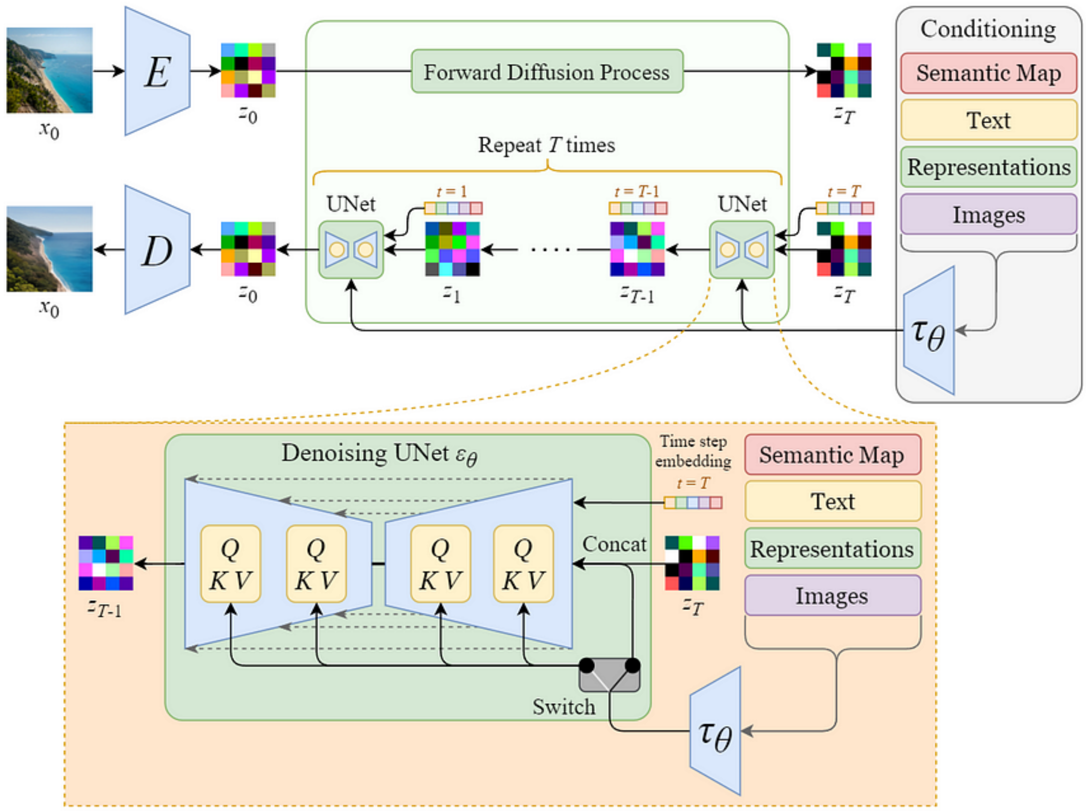


Figure 3.2: Stable Diffusion architecture [45]

(VAE), is the neural network used to compress high-resolution images into the latent space without losing significant information based on the manifold hypothesis in machine learning [48] .

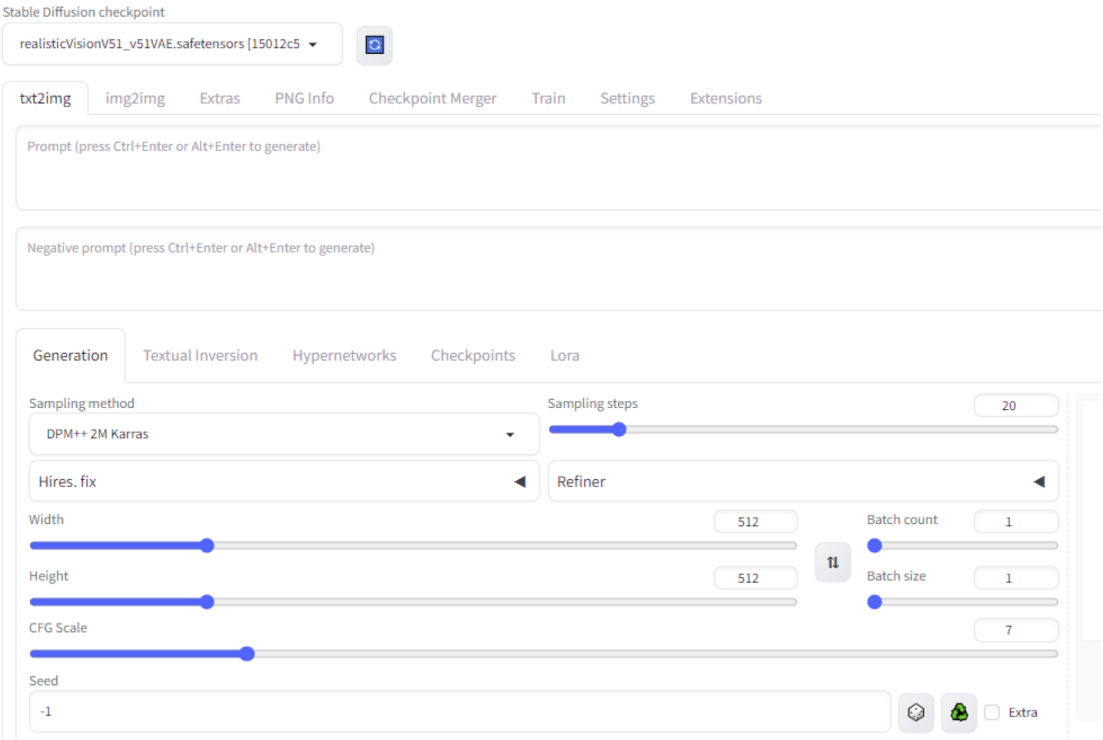
The VAE consists of two main components: an encoder E and a decoder D. The encoder compresses an image to a lower-dimensional representation in the latent space. The decoder restores the image from the latent space. The U-Net architecture is responsible for predicting the denoised image representation from the noisy latents, effectively subtracting predicted noise from the noisy latent image to refine the image representation.

Text prompts are transformed into embeddings by the CLIP tokenizer [49].
CLIP is a deep learning model developed by Open AI to produce text descriptions of any images. These embeddings are used to condition the noise predictor U-Net in the denoising process, ensuring that the generated images are not only visually coherent but also semantically aligned with the input text.

This approach is one of the preferred choice in AI-driven images generation since unlike its counterparts like DALL-E [50] and Midjourney [51], which rely on cloud services, Stable Diffusion operates efficiently on consumer-grade hardware, requiring only a GPU with a minimum of eight gigabytes of RAM.
As an open-source model, Stable Diffusion benefits from a community-driven development the model is continuously refined and expanded upon, with new applications and enhancements emerging regularly.
Offers users a high degree of control over the photorealistic generation process and the possibility to specify detailed prompts, leading to tailored and precise outputs.
While generative models are known for their element of randomness, Stable Diffusion reduces unpredictability in its outputs. This controlled randomness ensures more consistent and reliable results, essential for applications requiring a high degree of consistency.
And lastly, the architecture of Stable Diffusion is inherently scalable, allowing for the seamless integration of new models and checkpoints, adapting easily to new datasets and requirements. This scalability ensures that the model can handle a diverse range of image generation tasks.

3.1.2 AUTOMATIC1111 GUI

For this project Stable Diffusion from AUTOMATIC1111 [52] was downloaded and run locally, text-to-image feature and a version 1.5 were utilized. Figure 3.3 is a visual representation of the interface proposed to the user.



Stable Diffusion checkpoint: realisticVisionV51_v51VAE.safetensors [15012c5

Sampling method: DPM++ 2M Karras

Sampling steps: 20

Width: 512

Height: 512

Batch count: 1

Batch size: 1

CFG Scale: 7

Seed: -1

Figure 3.3: Stable Diffusion GUI

Some of the key settings of the model include:

Parameter	Description
Prompt	The primary creative input space where the user writes a textual description of the desired image. The detail in the prompt guides the SD generative process.
Negative Prompt	Allows specification of elements to be excluded from the generated image, refining the output by preventing certain themes or features.

Parameter	Description
Negative Prompt	Allows specification of elements to be excluded from the generated image, refining the output by preventing certain themes or features.
Sampling Steps	Determines the number of iterations for the denoising process, balancing between elaboration and computation time.
Sampling Method	Select the method for handling sampling.
Width & Height	Set the desired dimensions of the canvas size output.
Batch Count	Number of times the image generation pipeline is executed.
Batch Size	Controls the number of images generation each time the pipeline is run.
CFG Scale	The Classifier Free Guidance scale parameter controls how much the model should respect the prompt. Higher values result in closer adherence to the prompt, while lower values offer more creative freedom.
Seed	-1 for random generation, or specify a seed number for consistent results across different prompts.

Table 3.1: Description of Parameters for SD Generative Process [53].

3.1.3 Controlling Image Diffusion Models

To further condition the image generation process in Stable Diffusion, I choose a combination of fine-tuning and checkpoint models, each contributing uniquely to the final output.

Realistic Vision

To obtain photorealistic imagery the main Checkpoint Model used is Realistic Vision V5.1 [54], it operates on a stable diffusion framework and uses SD 1.5 as it's base model.
It is tailored to produce high-resolution, photorealistic images, so SD is conditioned to prioritize realistic textures and details, making it particularly effective for generating images that closely resemble real-world subjects.

The suggested prompt is:

```
RAW photo, subject, 8k uhd, dslr, soft lighting, high quality, film
    grain, Fujifilm XT3
```

While the negative prompt proposed is:

```
Deformed iris, deformed pupils, semi—realistic, cgi, 3d, render,
    sketch, cartoon, drawing, anime, mutated hands and fingers:1.4),
    (deformed, distorted, disfigured:1.3), poorly drawn, bad anatomy,
    wrong anatomy, extra limb, missing limb, floating limbs,
    disconnected limbs, mutation, mutated, ugly, disgusting,
    amputation.
```

LoRa

To protrait a specific subject the LoRA [55], Low-Rank Adaptation, models were involved, they are significantly smaller in size compared to full checkpoint models, usually ranging from 2 to 200 MBs, they can't be used alone but are designed work in conjuction with a base model checkpoint file and can apply subtle yet impactful changes.
LoRA models excel in generating high-quality images of particular styles or characters. They specifically modify the cross-attention layers of SD models, where the image and text prompts interact, allowing for customized AI art outputs without a substantial increase in storage requirements [56].

In this thesis, I directed the focus on training LoRA models towards specific subjects, using a starting dataset of approximately 80-100 images per person. This approach allowed for the generation of outputs that were both detailed and stylistically consistent. These models, adept at maintaining the subject's fidelity, also provided flexibility in altering backgrounds and lighting settings, ensuring a fine-tuned portrait. Once a LoRA model is trained, the files are placed in the "stable-diffusion-webui/models/Lora" directory and can be utilized for image generation. It is integrated in the prompt and can substitute the *subject* keyword, as previously mentioned in the Realistic Vision default prompt. The syntax for this integration is as follows:

```
<lora: name: weight>
```

The *name* refers to the specific LoRA model subject, and *weight* refers to the model's influence on the output. The default weight is set to 1, but this can be adjusted; setting it to 0 effectively disables the trained model's impact.

ControlNet

The ControlNet [57] neural network structure adds spatial conditioning controls to large, pretrained text-to-image diffusion models. It learns task-specific conditions and can effectively control SD model by enabling more conditional inputs in addition to the text prompt like edge maps, segmentation maps and human keypoints. In the SD model, the original weights remain frozen during the fine-tuning process and an external, trainable network is created specifically to handle the new conditional input. This network, operating outside the main SD model, is trained to inject additional information into the main model during the decoding phase. The key aspect of this setup is that the main model's architecture and weights don't get any alteration, ensuring its core functionalities remain intact.
As shown in Figure 3.4 to achieve specific poses for the output images the conditioning model used in combination with the ControlNet framework is OpenPose [58], a computer vision library for human pose estimation.

OpenPose detects 130 human keypoints, including eyes, nose, neck, shoulders, elbows, wrists, knees and ankles, from the driving images and uses them as an extra external condition for SD together with the text prompt.
In this project, I employed the *OpenPose_face* option, which is capable of extracting 70 facial keypoints in addition to the overall body position.

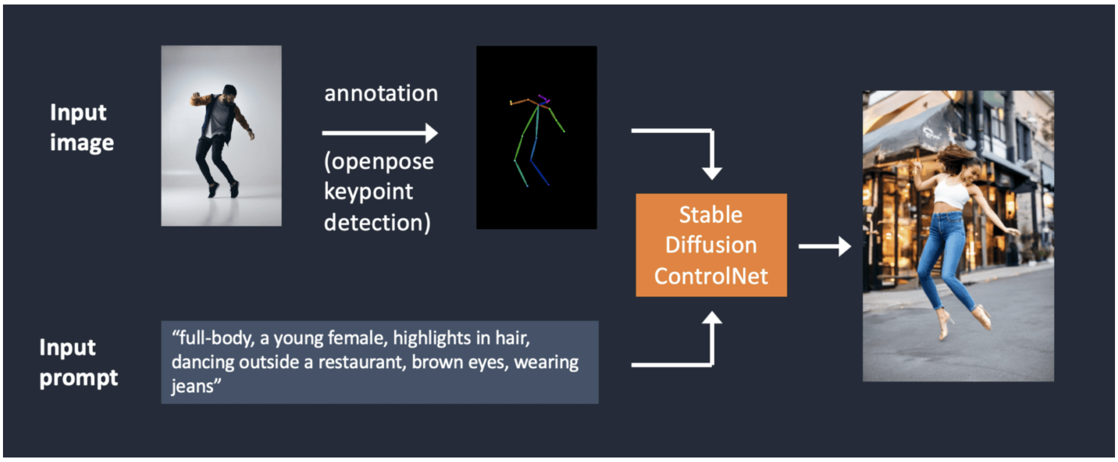

Input image
annotation
(openpose keypoint detection)
Stable Diffusion ControlNet
Input prompt
"full-body, a young female, highlights in hair, dancing outside a restaurant, brown eyes, wearing jeans"

Figure 3.4: ControlNet workflow using OpenPose [59]

3.2 Unreal Engine

For this thesis, I developed a Blueprint Widget [60] in Unreal Engine 5.2, Figure 3.5, using the platform's visual scripting system. It facilitates the positioning of MetaHumans and cameras within the project environment, mirroring the configuration used to generate 3D heads with FaceBuilder in Blender. The goal was to generate face portrait renders for two distinct applications: objective analysis, which involved cosine similarity computations, and subjective assessments conducted through surveys.

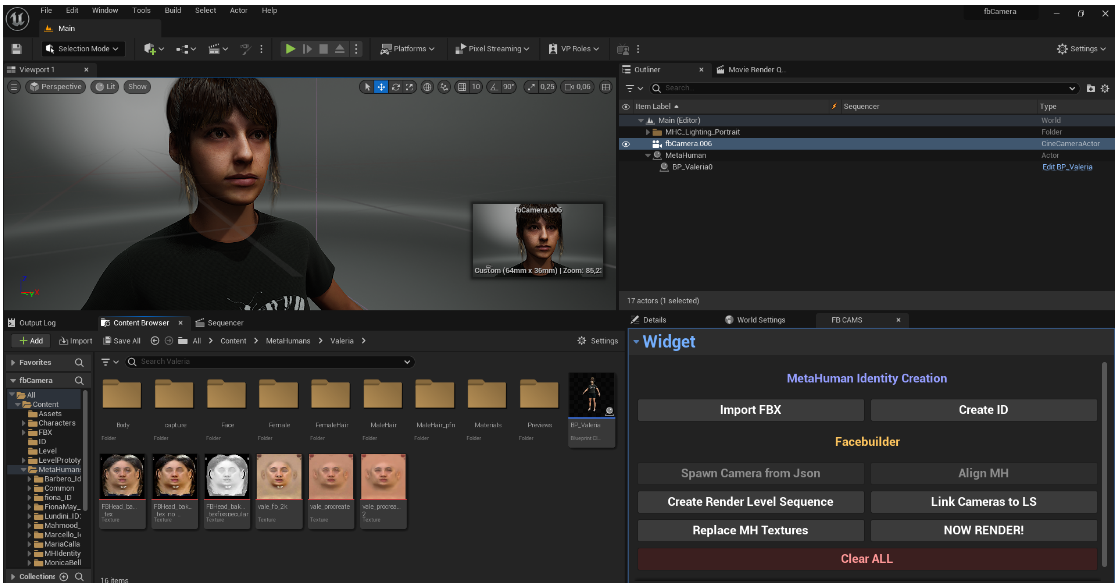


Figure 3.5: The widget interface within the Unreal Engine project

3.2.1 Widget Blueprint

A critical aspect of this development was aligning the differing parameters between the two software systems. For instance, I designed the **"Spawn Cameras from Json"** in Figure 3.6, button to parse JSON data that details camera configurations from Blender and instantiate them within Unreal Engine.

While Blender's default coordinate system has a positive Y-axis going into the screen, Unreal's Y-axis extends out of the screen. Similarly, camera parameters needed conversion, such as translating resolution from pixels in Blender to millimeters in Unreal, and focal length from degrees to millimeters.

The orientation of the camera also required adjustment:

- Blender's pitch (rotation around the X-axis) corresponds to the Y-axis rotation in Unreal.

- Blender’s yaw (rotation around the Z-axis) matches the Z-axis rotation in Unreal.

- Blender’s roll (rotation around the Y-axis) aligns with the X-axis rotation in Unreal.

By default, cameras are created facing the positive X-axis in Unreal, whereas in Blender, they face the positive Y-axis.

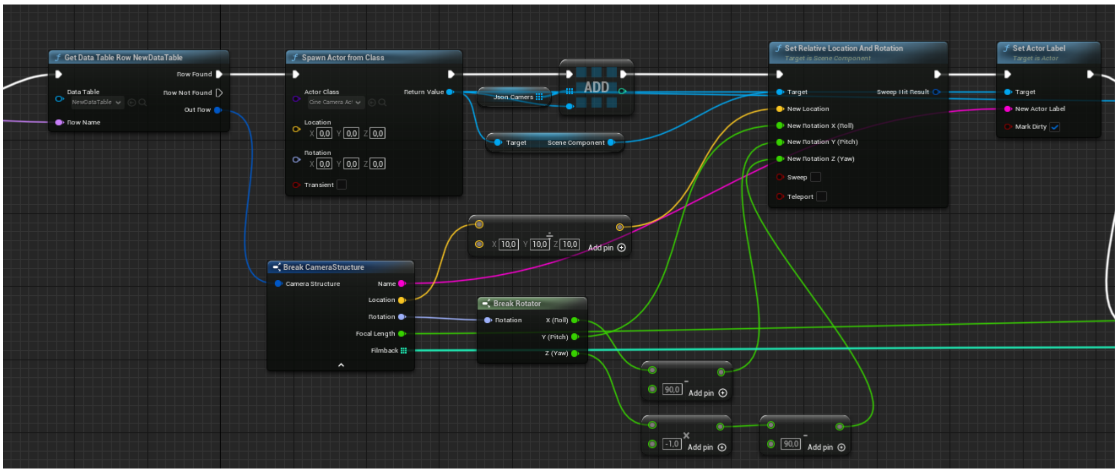


Figure 3.6: A snippet of the "Spawn Cameras from Json" functionality

The **"Create Render Level Sequence"** button in the widget is designed to automate the process of generating a render sequence within Unreal Engine. It compiles all the automatically positioned cameras in the scene into a Level Sequence, setting up each camera view to correspond with a single frame in the timeline. This allows for batch rendering of multiple images of the MetaHuman in various poses without the need to manually track or set each camera to last for just one frame.

The **"Link Cameras to LS"** in Figure 3.7, is a function that binds the cameras to the Level Sequence, ensuring that each camera’s view is correctly represented in the sequence. This setup streamlines the workflow, enabling a quick and efficient way to render out multiple angles or poses of a character with the **"Now Render"** button, which selects the images’ output settings and triggers the rendering process of the assembled sequence.

The **"Align MH"** button assists in the precise positioning of a MetaHuman asset in Unreal Engine’s scene.
It allows the user to select the MetaHuman asset and automates the placement

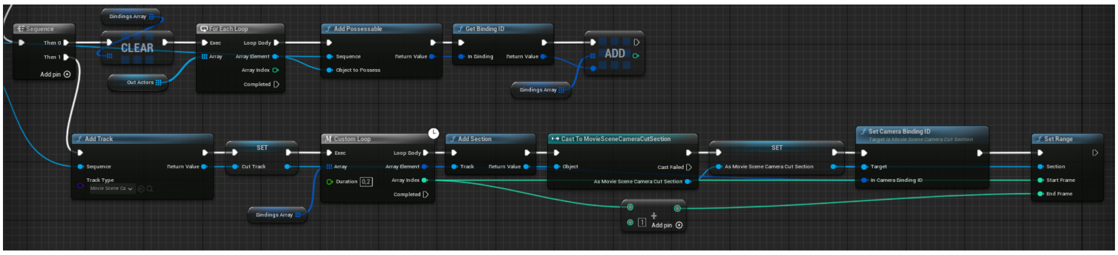


Figure 3.7: A snippet of the "Link Cameras to LS" functionality

process by automatically centering the selected MetaHuman from the content browser onto the scene's origin. Additionally, it adjusts the asset's pivot point from the default feet location to the center of mass of the head, facilitating easier manipulation and placement of the character within the scene environment. This tool is particularly useful for ensuring that the MetaHuman is correctly oriented for subsequent tasks such as animation or rendering.

The **"Replace MH Textures"** in Figure 3.8, is a function that simplifies the process of updating the facial appearance of the selected MetaHuman by ensuring that all relevant facial textures are replaced in a single action. After selecting the MH whose facial texture needs updating, and then choosing the new texture file in the content browser, the button triggers a process within the blueprint which systematically updates the facial material's textures with the selected file. This utility ensures that all associated textures are cohesively updated, thereby significantly simplifying the character customization for the rendering process.

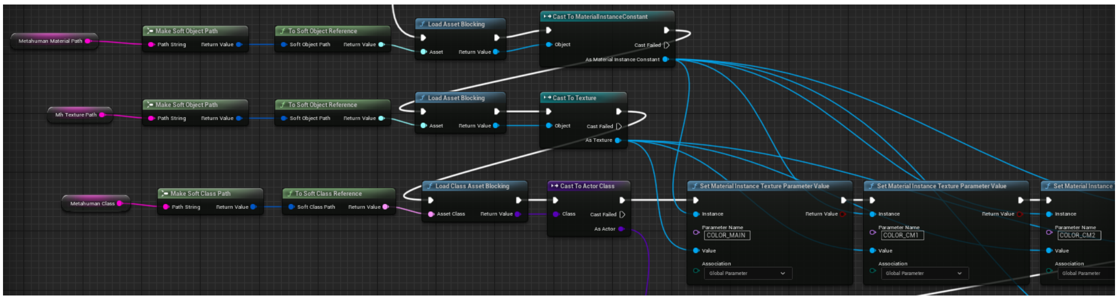


Figure 3.8: A snippet of the "Replace MH Textures" functionality

Chapter 4

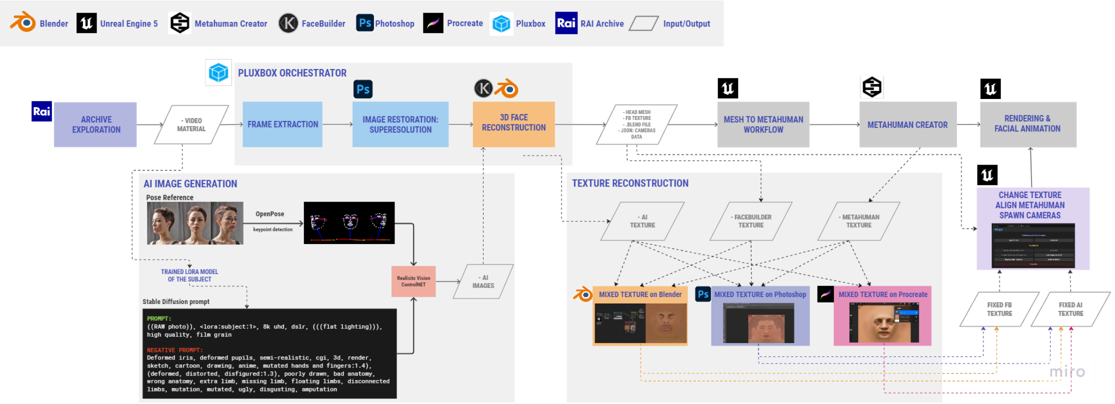
Methodology

4.1 Design and Implementation of Workflow

This chapter discusses the practical application of the technologies previously introduced, Figure 4.1 illustrates the proposed enriched workflow for synthetic human creation.

The process initiates with an extensive collection of web and RAI's archive images, for each subject, the most relevant ones are then used for the 3D face reconstruction phase, which focuses on the generation of a 3D model that closely approximates the real subject head.

The next phase explores various methods to create comprehensive and model-compatible facial textures. In addition to traditional techniques, experimentation with generative AI techniques has produced promising results in addition to standard methods.



```
PROMPT:
((RAW photo)), <lora:subject:1>, 8k uhd, dslr, (((flat lighting))),
high quality, film grain

NEGATIVE PROMPT:
Deformed iris, deformed pupils, semi-realistic, cgi, 3d, render,
sketch, cartoon, drawing, anime, mutated hands and fingers:1.4),
(deformed, distorted, disfigured:1.3), poorly drawn, bad anatomy,
wrong anatomy, extra limb, missing limb, floating limbs, disconnected
limbs, mutation, mutated, ugly, disgusting, amputation
```

Figure 4.1: Implemented Workflow for Synthetic Human Creation.

The final phase of the workflow introduces automation in the posing and rendering tasks for Metahumans. This simplification facilitates the creation synthetic human evaluation material used in the following chapter.

4.1.1 Dataset Compilation

In the collection of the dataset, nine subjects were considered, and although a complete set of 15 images per subject would be sufficient for demonstration purposes, the goal was to achieve the most detailed results possible. Therefore, a decision was made to collect around 80 to 100 images for each subject, influenced by discussions in forums and recent YouTube videos.

The most relevant images, approximately ten, that capture the distinct facial details were utilized for the 3D head model construction using Facebuilder.

For the LoRA training of a person or a character model, it is generally advised to use a varied dataset, typically ranging from 30 to 100 images, with a substantial number being close-ups and body shots to capture the subject's features.

The versatility of the dataset is crucial, as the training outcomes are robust, showing that the variability in the poses, outfits, and backgrounds, including different zoom angles and illumination levels, does not affect the training process, as shown in Figure 4.2.

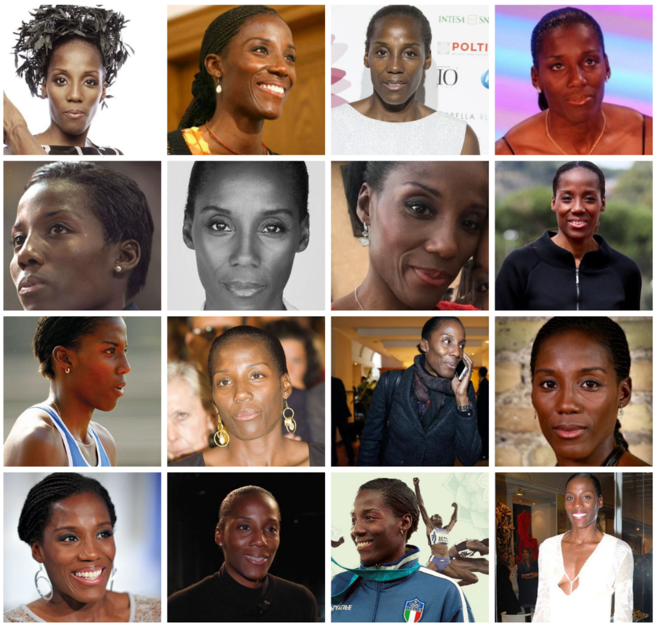
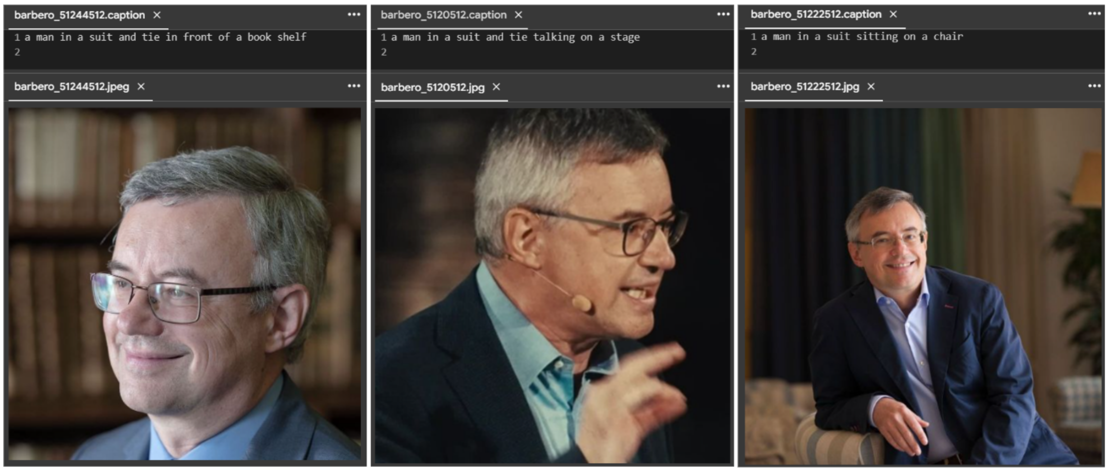


Figure 4.2: Part of the Fiona May's dataset.

This flexibility is due to the advanced capabilities of the training models, which can handle a wide spectrum of image qualities, from low to high resolution. Supported file formats include .png, .jpg, .jpeg, .webp, and .bmp, showcasing the system's adaptability to different data types.

4.1.2 Dataset Optimization for LoRA training

During the pre-processing phase of the training, all images were cropped to a uniform size of 512x512 pixels, focusing on the face, as the primary intent was to generate AI images that detailed the face more than the body and also because the training process expects the images with the stated dimension. This level of detail is essential since the face carries significant individual characteristics, which are fundamental for creating a realistic Synthetic Human.
BLIP captioning, a CLIP model described in the article "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation" [61] was employed to enhance the dataset with rich descriptive data, providing a framework that creates accurate captions that are used into the training of the LoRA model, as shown in Figure 4.3.



```
a man in a suit and tie in front of a book shelf
```

```
a man in a suit and tie talking on a stage
```

```
a man in a suit sitting on a chair
```

Figure 4.3: Examples of BLIP captioning on Alessandro Barbero's images.

4.1.3 Effective Prompts for Texture Synthesis in Stable Diffusion

The construction of the prompt is a crucial element in guiding the customized Stable Diffusion model to produce the desired output.
Let's explore how specific syntax within a prompt can direct the SD's attention

and influence the generated image.

Attention/Emphasis Modifiers: Modifiers such as parentheses and square brackets can adjust the model's focus on certain aspects of the prompt. Using parentheses around a word or phrase, like ($bright$), increases the model's attention on the concept of brightness, making it a prominent feature in the resulting image. Conversely, square brackets, [$blurry$], signal the model to de-emphasize that attribute, leading to less focus on blurriness.

Specifying Weights: Weights can be assigned to specific terms within the modifiers to fine-tune their influence. For example, ($text$: 1.4) applies a weight that increases the term's emphasis beyond the default one. If no weight is specified, it is treated as if it has a weight of 1.1. Similarly, using weights in square brackets diminishes the focus, such as [$text$: 0.9], which would reduce the attention given to the term "text." The aim of the chosen prompt is to condition Stable Diffusion to generate a high-detail image that looks like a non-processed photograph of a subject (with characteristics defined in the LoRA-trained model), under even lighting, against a white background, and with the subject having a neutral facial expression.

Modifier	Effect
`(word)`	Increase attention to word by a factor of 1.1
`((word))`	Increase attention to word by a factor of 1.21 (1.1 * 1.1)
`[word]`	Decrease attention to word by a factor of 1.1
`(word:1.5)`	Increase attention to word by a factor of 1.5
`(word:0.25)`	Decrease attention to word by a factor of 4 (1 / 0.25)

Table 4.1: Modifiers and weights and their effects on attention in prompts from AUTOMATIC1111's Wiki. [62]

```
RAW photo, <lora:subject:1>, ((neutral expression)), (((flat
    lighting))), white background
```

- **RAW photo:** Indicates that the output image should have the qualities of a RAW photograph, meaning it should be unprocessed and contain all the original data for high-quality editing.
- **<lora:subject:1>:** Specifies that the image should be guided by a subject-specific model fine-tuned to a high degree, denoted by :1, significantly affecting the features, style, or identity of the subject.

- **((neutral expression)):** The double parentheses emphasize the importance of the subject having a neutral expression, crucial for a realistic representation without any emotional connotation.
- **(((flat lighting))):** Strong emphasis on flat lighting, suggesting an even, diffused light that minimizes shadows and highlights to clearly show the subject's features.
- **white background:** Specifies a white background for its simplicity and to ensure that the subject stands out without any background details influencing the subject's perception.

For some generations the prompt has been tailored to achieve specific results, like for female subject, it was necessary to specify hairstyles like *"slick back ponytail"* or *"bald"* to prevent hair from obscuring facial features, ensuring a clear view of the face.
When dealing with a diverse database, specifying the age with *"of X years old"* helped in achieving a consistent look across different subjects by matching the age-related features more accurately. In Figure 4.4 an example of hair and age-tuning. Asking for specific outfits ensured that the clothing did not interfere with the portrayal of the subject's features, like covering the neck with scarves or jackets,

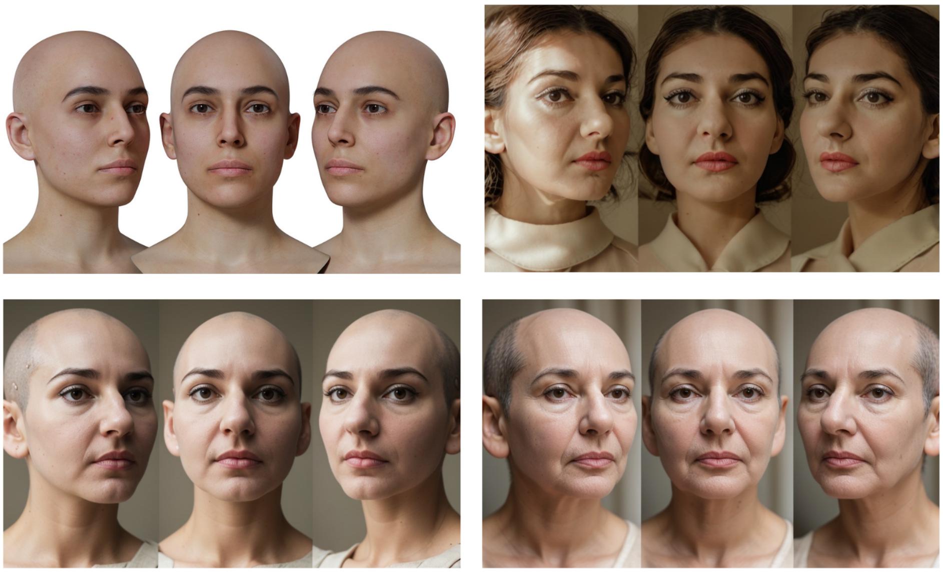


Figure 4.4: Starting from the top left, the driving image used for Controlnet, then a representation of 20 years of age Maria Callas, followed by her at 35 years, lastly her at 70 years old.

which is important for a uniform appearance, an example illustrated in Figure 4.5. For some subjects, it was necessary to alter the weight given to certain keywords, like *"neutral expression"*, to either tone down or accentuate specific facial expressions for consistency across all images.

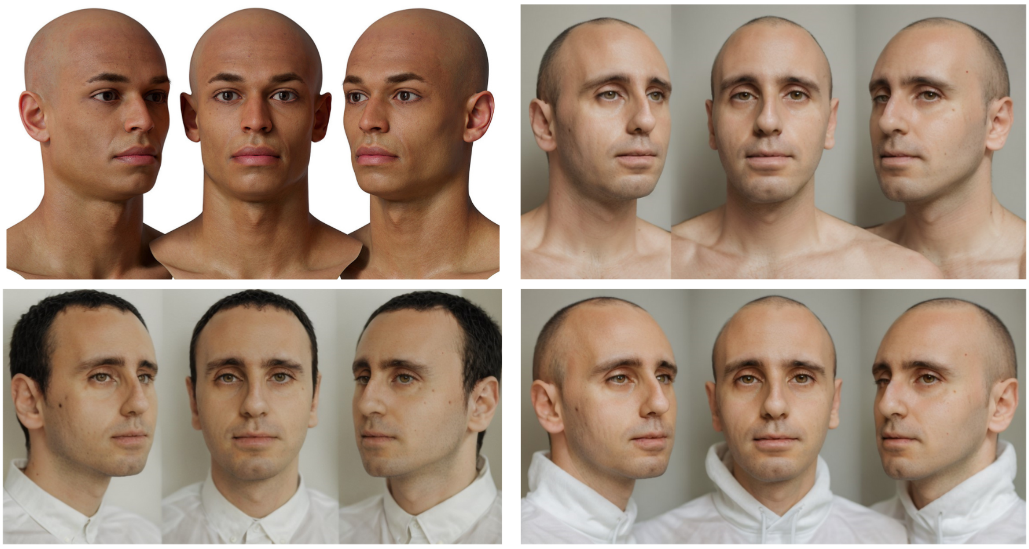


Figure 4.5: Starting from the top left, the driving image used for Controlnet, then a series of clothes and hair tuning on Zerocalcare.

The negative prompt employed follows the guidelines outlined in the previous chapter, adhering to the suggested Realistic Vision parameters, but also in this case some keywords were used strategically to exclude unwanted elements such as earrings, necklaces, glasses, or any accessories that could cover the face, such as in Figure 4.6, but also to remove any potential watermarks coming from the images of the LoRA training set.
The advantage of using RealisticVision which is trained for realistic subjects, is that it was not required to specify in the negative prompts to avoid non-realistic styles like cartoons, anime, 3d, or paintings.
Also the LoRA model itself was sufficiently trained to understand the context of realistic human representation without the need for such explicit instructions.

4.1.4 Subject Posing

Although ControlNet and OpenPose settings were not included directly in the text prompt, they played a crucial part through a separate section in the Stable

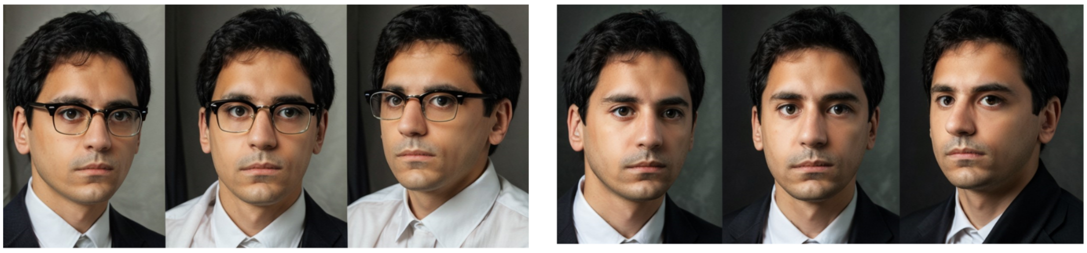
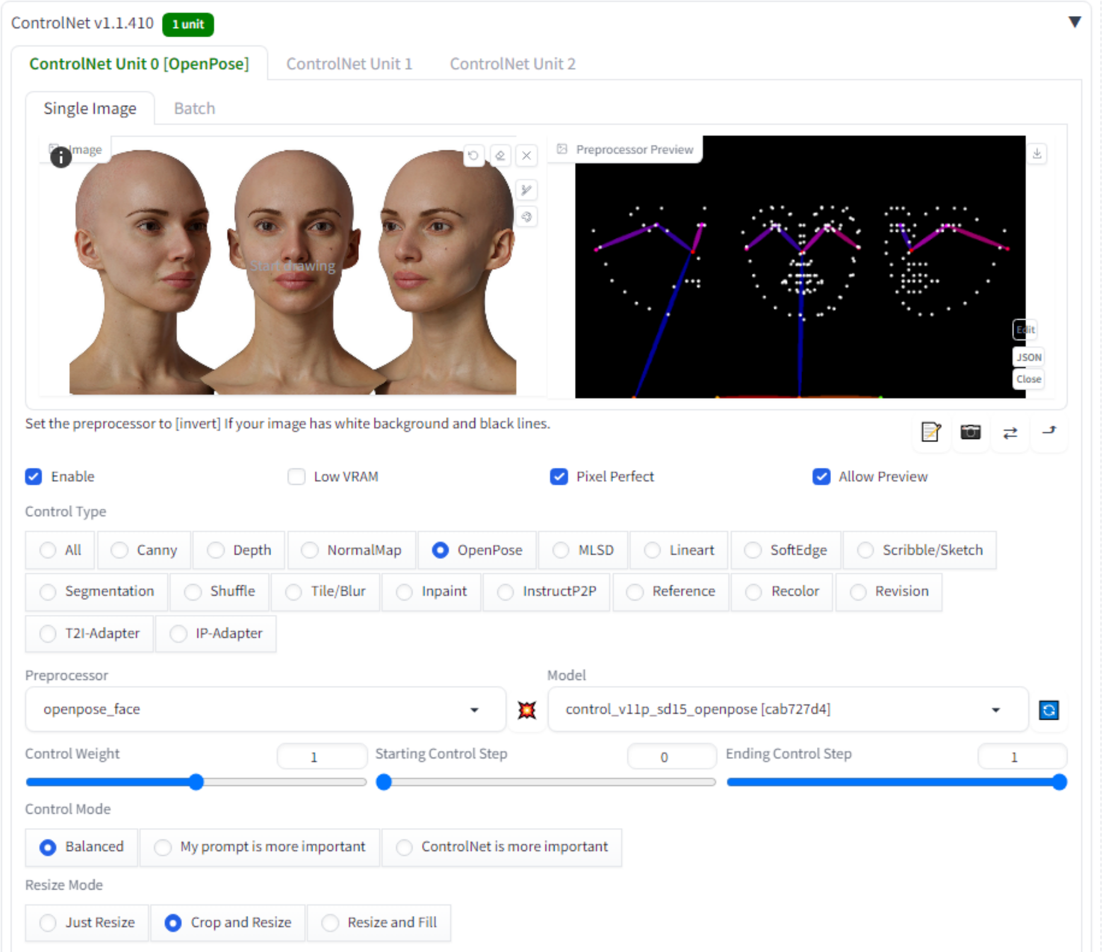


Figure 4.6: An Example of eyeglasses removal on Valerio Lundini.

Diffusion interface.
The *"Balanced"* option was chosen, as in Figure 4.7, to ensure a fair compromise



ControlNet v1.1.410 1 unit

ControlNet Unit 0 [OpenPose] | ControlNet Unit 1 | ControlNet Unit 2

Single Image | Batch

Image | Preprocessor Preview

Set the preprocessor to [invert] If your image has white background and black lines.

☑ Enable ☐ Low VRAM ☑ Pixel Perfect ☑ Allow Preview

Control Type

○ All ○ Canny ○ Depth ○ NormalMap ● OpenPose ○ MLSD ○ Lineart ○ SoftEdge ○ Scribble/Sketch ○ Segmentation ○ Shuffle ○ Tile/Blur ○ Inpaint ○ InstructP2P ○ Reference ○ Recolor ○ Revision ○ T2I-Adapter ○ IP-Adapter

Preprocessor: openpose_face

Model: control_v11p_sd15_openpose [cab727d4]

Control Weight: 1 | Starting Control Step: 0 | Ending Control Step: 1

Control Mode

● Balanced ○ My prompt is more important ○ ControlNet is more important

Resize Mode

○ Just Resize ● Crop and Resize ○ Resize and Fill

Figure 4.7: The ControlNet Openpose section in Stable Diffusion UI.

between the text prompt and ControlNet's input, allowing for the AI-generated image to maintain equal fidelity to both the text prompt and the pose indicated by the keypoints extracted from a selected reference image.
This option ensures that the key facial features and expressions match the specifications of the prompt, while the poses accurately reflect those of the driving image. The inspiration for the poses was drawn from professional subject scanning photographs, which are typically utilized for high-resolution facial scans.
The driving images for the ControlNet were sourced from 3D Scan Store [63], which offers a variety of ultra high-definition 3D head scans suitable for detailed and accurate modeling.

Professional facial scanning practices, often involve capturing a wide range of angles through numerous photographs to create a detailed facial scan, in fact, they include a 90-degree profile, the back of the head, and various neck tilt positions. For the specific needs of this project and the texture mapping phase, the generated AI images focus only on three poses: directly front-facing and 45 degrees to each side.

It's recognized that achieving perfect consistency in character profiles can be a significant challenge in text-to-image generation, particularly when the training set lacks a diverse range of profile shots.
In this thesis, less known subjects were intentionally selected, that may not have as extensive range of images or photoshoots as more publicized figures or celebrities, and this was done to mainly test the robustness of the model.
It's important to note, though, that exact replication of the subject's head shape is not crucial in this context since the primary objective is to extract detailed facial textures using tools like FaceBuilder under consistent lighting conditions.
Therefore, even if the profiles do not perfectly match the original subjects, the quality and intricacy of the texturing remain high, meeting the project's requirements for this phase of synthetic human creation.

4.1.5 3D Head Reconstruction

Following the collection of web images for each subject, and the generation of the AI ones, the workflow transitions to the 3D head reconstruction phase where a plugin named "Export cameras info" has been developed.
This plugin executes the **FB_Head_Reconstruction_Automatized.py** script, which is critical to this workflow. The script is based on FaceBuilder's local capabilities to automate the 3D face reconstruction process as elaborated in Chapter 2. When executed it performs tasks typically handled by FaceBuilder in a single step, such as "Automatic Camera Setup" which reads metadata from photos to configure virtual cameras automatically and "AI-powered Automatic Face Alignment" which

eliminates the need for manual pinpointing of facial features to shape the model. The output from this process is comprehensive, including an fbx file with the 3D head mesh, the extracted texture from the images, a Blender scene for any additional manual mesh adjustments, and a JSON file with camera information.

In addition to the local version, the "FB_Head_Reconstruction_Automatized.py" script has been adapted for use within the **Pluxbox Orchestrator**.
Pluxbox [64] offers a no-code development Orchestrator to assist users to create comprehensive and customized media and software management system.
For this project, it offers an interface to identify and extract significant frames of the subject from RAI's video archives. These frames are then processed to apply super-resolution using Adobe Photoshop's API, enhancing the image quality before they serve as input for 3D reconstruction. Subsequently, the Orchestrator, through the "FaceG3n" service, invokes the script to generate the 3D head mesh. This process showcases the fusion of media and software management by showing the transformation of historical footage into high-fidelity digital assets.

4.1.6 Texture Reconstruction

After creating the Metahuman using the "Mesh to Metahuman" workflow in Unreal Engine, it is possible to export the "Metahuman texture", that will be referred as **Original_MH**. In conjunction with this, two additional textures are obtained from the preceding processes: the "FB texture" and the "AI texture".
The "FB texture", referred as **Original_FB**, is derived from the web-collected images processed by the "Export cameras info" plugin, which is part of the FaceBuilder tool suite. This texture captures the authentic details of the subject's face as depicted in the source photographs.
The "AI texture", called **Original_AI** on the other hand, is generated from images created by Stable Diffusion. This AI-driven process synthesizes textures that can potentially reconstruct missing features or enhance existing ones.
The three textures presented; *MH_Original, FB_Original, and AI_Original*, serve as base elements for the texture reconstruction phase. Given that these textures are not fully refined, they require further enhancement to achieve the desired quality and realism when applied to the Metahuman model.

To accomplish this, three main techniques to blend the textures together are employed:

- **Photoshop:** Involves the manual process of layering and merging textures. Photoshop's comprehensive toolset allows for precise control over how textures are combined. This method is hands-on, requiring user input to adjust the layer blending modes and opacity to achieve the seamless integration of textures.

- **Procreate:** Uses the app's touch interface and advanced brush system, which, when used with a stylus on a tablet, gives a tactile, hands-on approach to texture refinement. This method allows for a more organic and nuanced application of details to the textures.
- **Blender:** This blending is semi-automated and employs Blender's node-based compositing system to combine textures. As shown in Figure 4.8, the goal is to mix the FB_Original and AI_Original textures with the MH_Original one to create refined textures for the Metahuman model.
  Initially, manual adjustments were made to the mask layer to ensure the textures align correctly with the Metahuman's facial features.
  The masking input was customized to feature a more pronounced white in areas that define recognizability—like the nose, lips, and eye contours—while fading to black in areas prone to artifacts or less visibility, such as nostrils, ears, and eyelids. This mask essentially functions as a map: the white areas dictate the presence of either the FB or AI texture, while the black controls the prominence of the MH texture.
  The ColorRamp node, plays a key role in determining the visual presence of either texture in the final blend. By manually adjusting the ColorRamp parameters or by introducing additional handlers it is possible to gain even more precise control over the blending process and the contribution of each texture can be fine-tuned.
  Once these settings are configured, the composited texture is quickly rendered, resulting in a tailored texture that suits the Metahuman model.

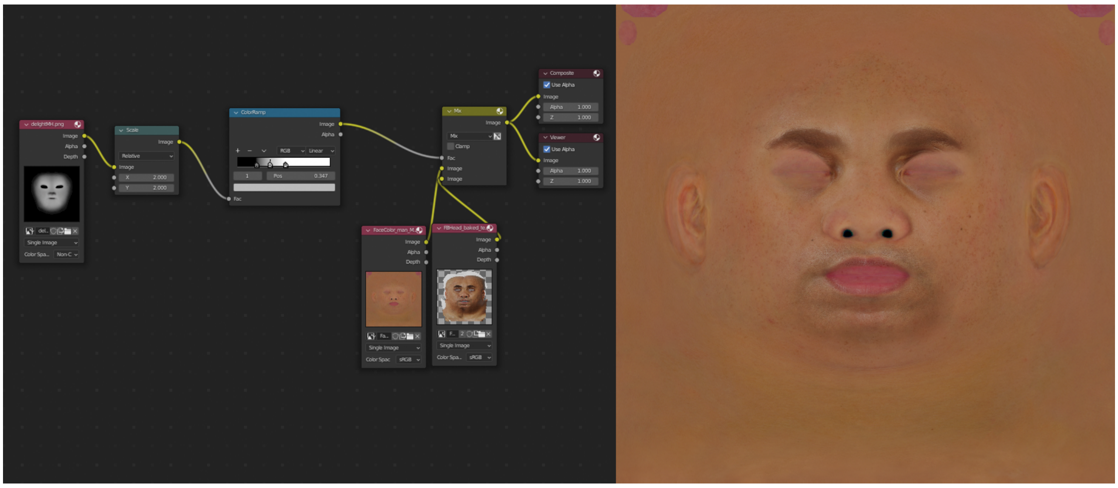


Figure 4.8: Blender Texture Blending and Compositing

4.1.7 Final Model Posing

The final segment of the workflow, as detailed in the previous chapter's 3.2.1 Widget Blueprint section, concerns the efficient and modular posing of Metahumans within the Unreal Engine scene, facilitating rapid rendering with the various produced textures to create material for the workflow's validation phase.

To address Unreal Engine's lack of a static scene origin, where assets typically appear near the cursor when dropped in to the scene, Metahumans are anchored at a designated point which all camera perspectives converge. Such placement is also important for centering on facial features, which will be essential during the facial animation phase.

Camera angles within the scene are meticulously aligned to match the perspective of those in Blender, their set up in the scene is based on a JSON file extracted from FaceBuilder, and parsed by the Widget, detailing each camera's location, rotation, focal length, and filmback dimensions, as shown below. This integration provides a coherent and unified visual continuity between the two software environments.

```
{
    "Name": "fbCamera",
    "Location": {
        "X": -23.405617475509644,
        "Y": 1212.1780395507812,
        "Z": 147.58141040802002
    },
    "Rotation": {
        "Pitch": -0.6774466565773083,
        "Yaw": -0.8986029525387379,
        "Roll": 83.55189411663594
    },
    "FocalLength": 98.7021713256836,
    "Filmback": [
        1920,
        1080
    ]
},
{
    "Name": "fbCamera.001",
    "Location": { ...
```

The examples in Figure 4.9 demonstrate this precise synchronization of camera positions with their Blender counterparts.

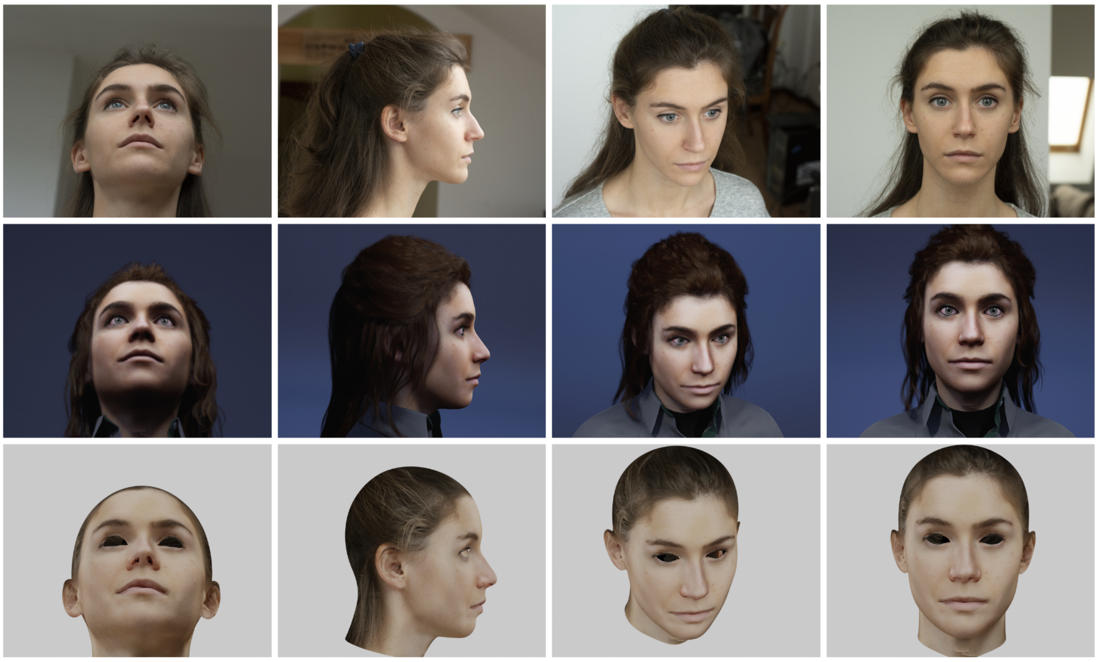


Figure 4.9: From the top row: The original images of the subject, Metahuman rendering on Unreal using the Widget, 3D Head mesh generated on Blender within FaceBuilder plugin.

Chapter 5

Model Validation

In this chapter, an evaluation is conducted to determine the reliability and robustness of the automated workflow proposed for generating Synthetic Humans.

The analysis will focus on the fidelity and similarity of the Synthetic Humans to the reference subjects. The evaluation will be conducted using both objective and subjective methods by comparing the original images depicting the subjects with the resulting Synthetic Humans.

5.1 Test Material Preparation

5.1.1 Representative Dataset

To conduct a thorough and meaningful validation of the model, it was crucial to define a representative dataset of Synthetic Humans.
This dataset was designed to encompass a variety of individuals representing different demographic characteristics, including varying ethnicities, ages, and genders. Additionally, for each subject, a collection of 80 to 100 pictures was assembled, depicting them in a range of poses and expressions, from close-ups to distant shots, without strict limitations on the variety of images.
Emphasis was given to Italian subjects, to create a more tailored dataset that aligns with the project's specific goals. The images were sourced from a blend of RAI's archive and online resources, they include a variety of historical periods, resulting in a diverse range of photographs from black and white, lower resolution images to contemporary, high-resolution ones.
The diversification of these features aims to ensure and demonstrate that the workflow is capable of generating Synthetic Humans reflecting the traits of a broad spectrum of individuals in the global population.

In Figure 5.1, photos depicting the subjects selected for the creation of the Synthetic Humans for the dataset are shown.

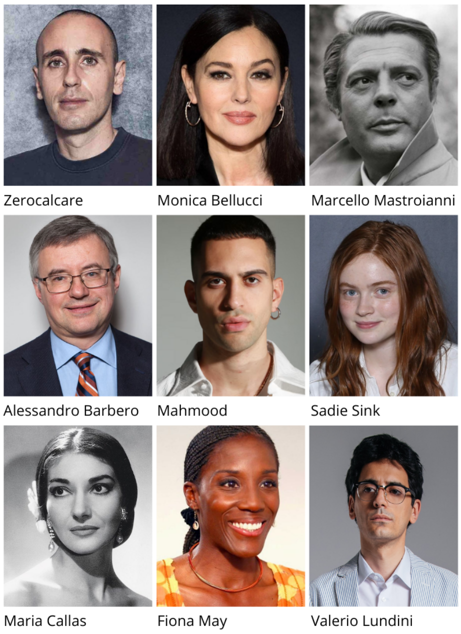
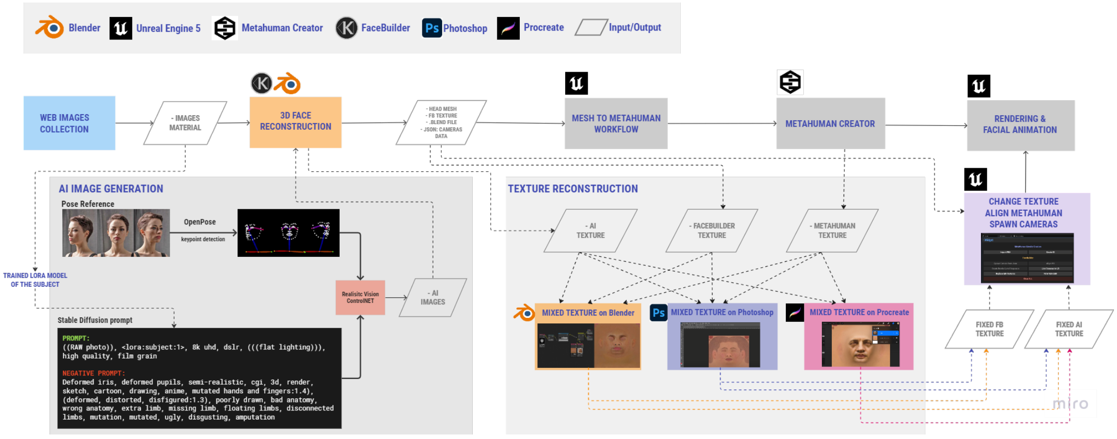


Figure 5.1: The selected subjects to which the workflow has been applied.

For each individual, the following phases were implemented as schematically illustrated in Figure 5.2.



```
PROMPT:
((RAW photo)), <lora:subject:1>, 8k uhd, dslr, (((flat lighting))),
high quality, film grain

NEGATIVE PROMPT:
Deformed iris, deformed pupils, semi-realistic, cgi, 3d, render,
sketch, cartoon, drawing, anime, mutated hands and fingers:1.4),
(deformed, distorted, disfigured:1.3), poorly drawn, bad anatomy,
wrong anatomy, extra limb, missing limb, floating limbs, disconnected
limbs, mutation, mutated, ugly, disgusting, amputation
```

Figure 5.2: The workflow followed

Initially, a set of approximately ninety images depicting the subject of interest was collected. These images were utilized in training a LoRA model designed to generate images of the individual using generative networks. A portion of the images collected for LoRA model training was allocated to the **Face Reconstruction** phase.

It is important to note that, for the generation of Synthetic Humans in this context, the orchestrator developed in collaboration with PluxBox was not employed. Instead, the script named ***"FB_Head_Reconstruction_Automatized.py"***, the same script used in the orchestrator via Rest API, was locally used to automate the creation of the **3D head mesh**. This script automatically generates the 3D model of the individual's face, utilizing the subject's images.

The output of this process includes an ***fbx*** file containing the 3D head mesh, the extracted texture from the images used to build the 3D model, the **Blender scene** (in case additional manual modifications to the mesh are desired), and a **JSON file** containing camera information. The latter is crucial for replicating the same view in Unreal Engine, facilitating the generation of renders of Synthetic Humans and subsequent model validation.

After obtaining the 3D head model from the original images, the same script was used with the images generated by the Generative AI. From the resulting output, only the texture is extracted, which will be used in the testing phase for model validation.

The availability of the 3D head mesh and textures generated by FaceBuilder facilitated the workflow, which includes importing the FBX mesh into Unreal and executing the ***"Mesh To Metahuman"*** procedure. This process utilizes the head mesh obtained during the Face Reconstruction phase as a base for creating a Metahuman.

After obtaining the MH, several facial textures were generated through a series of blending operations. The base textures, that can see in figure 5.3, employed in these operations are as follows:

- **Original MH**: The base texture provided by the Metahuman framework, designed for high compatibility with its models.
- **Original FB**: Textures obtained from FaceBuilder mapping using the subject's web original images.
- **Original AI**: Textures generated by FaceBuilder mapping using the subject's Stable diffusion generated images.

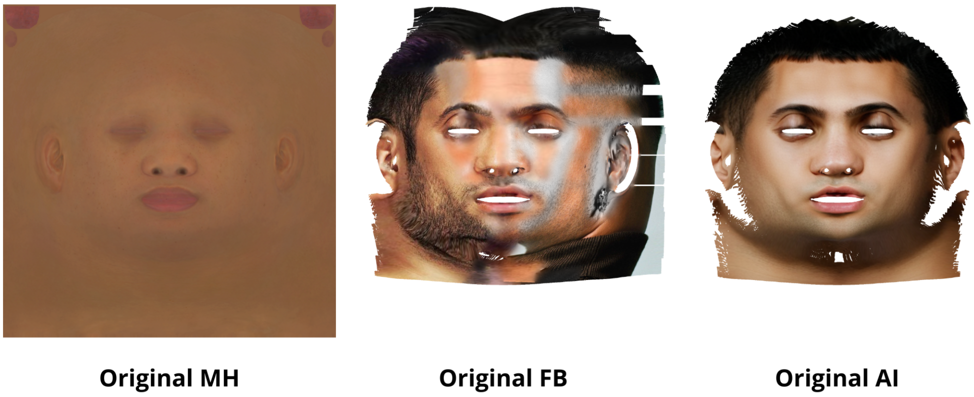


Figure 5.3: The base textures used for the various blending methods

The blending process involved combining the base textures to address any gaps and ensure a cohesive appearance. The Original MH texture was essential in this process as it helped fill in any missing details and provided a consistent base for the blend. The following table 5.1 provides an overview of the blending methods and the specific textures that were combined.

Textures Generated from Original Images			
Name	**Blending Method**	**Texture A**	**Texture B**
Original_Blender_Mix	Blender	Original FB	Original MH
Original_Photoshop_Mix	Photoshop	Original FB	Original MH
Textures Generated from AI Images			
Name	**Blending Method**	**Texture A**	**Texture B**
Procreate_AI	Procreate	Original MH	Original AI
Blender_AI	Blender	Original MH	Original AI
Photoshop_AI	Photoshop	Original MH	Original AI

Table 5.1: Texture blending methods overview

During the testing phase, all obtained textures were used to objectively and subjectively evaluate which texture generation procedure is more effective in ensuring similarity with the target subject.

The blended textures were imported into Unreal Engine 5, and placed in the corresponding Metahuman project folder.
Through the implementation of a **Widget** in Unreal several processes were automated: loading of the Metahuman into the scene, instantiating cameras to match the views from the Face Reconstruction images (guided by the JSON file), and swapping textures on the Metahuman.
Consequently, this automation facilitated the Metahuman's positioning in the same poses used during the Face Reconstruction phase, allowing for a direct comparison between the original image and the Metahuman rendering.

5.1.2 Evaluation Metrics

Validating our Synthetic Humans generation model required a careful approach to identify metrics that would reflect human perception. A significant part of our work was dedicated to studying and understanding which metrics could effectively capture the human experience in evaluating generated faces.

In addition to objective metrics, we conducted a subjective validation through a **survey** conducted on a sample of **46 participants**. This allowed us to gain a human perspective on facial resemblance and to support our choice of objective metrics.

Survey: Subjective Evaluation

To obtain an assessment reflecting human perception of the generated Synthetic Humans and the intermediate stages used to achieve our goal, we conducted a survey with a significant number of participants. The survey questions were strategically formulated to mirror the same aspects examined through objective metrics. This approach provided us with a complementary and human perspective on the fidelity of synthetic faces.

Analysis of Objective Metrics

In the model validation process to measure the effectiveness of the system in generating Synthetic Humans, we focused on two objective metrics: the **Cosine Similarity** and the **Euclidean distance**.

Cosine Similarity Metric

The choice of the **Cosine Similarity metric** [65] is based on its ability to measure the similarity between high-dimensional vectors, making it particularly suitable for complex data such as facial images. The cosine similarity metric is a method for calculating the similarity between vectors based on the angle between them in the vector space. The calculation is performed using the cosine formula:

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$$

This formula calculates the angle (θ) between two vectors ($\boldsymbol{a}$ and $\boldsymbol{b}$) and returns a value ranging from -1 to 1. A value of -1 indicates that the vectors point in opposite directions, while a value of 1 indicates that the vectors are perfectly aligned. A value of 0 indicates independence between the vectors.

This metric is widely used in the field of facial recognition because it reflects how humans perceive the similarity between faces. When two faces are very similar, we expect their cosine similarity score to approach 1, making it an ideal tool for evaluating the similarity between subjects' faces and Synthetic Humans.
Furthermore, the robustness of the cosine similarity metric to variations in lighting conditions and different facial poses ensures that small changes in the subject's appearance or pose do not significantly affect the score.

One of the most significant applications of the cosine similarity metric is through the **ArcFace** algorithm.

ArcFace: Maximizing Cosine Similarity

ArcFace [66] is an advanced deep learning-based facial recognition algorithm. Its distinctive feature is the use of the cosine metric to calculate the similarity between features extracted from faces. During the training phase, ArcFace is configured to optimize a specific loss function based on the cosine metric.
ArcFace leverages a loss function called "Cosine Margin Product", designed to maximize the similarity between feature vectors of faces from the same individual (intra-class) and minimize the similarity between feature vectors of different individuals (inter-class). In simple terms, the algorithm calculates the cosine angle between the feature vector extracted from a facial image and vectors representing class centers.
During training, the model learns to position feature vectors of faces from the same person closer to each other in the feature space.
ArcFace is one of the most advanced facial recognition models due to its ability to robust face detection, recognition, and landmark detection capabilities.

InsightFace Library

ArcFace within the InsightFace library [67] is utilized to extract the "landmark_2d_106" array, which consists of 106 pairs of (x,y) coordinates.
Each pair represents the position of a specific facial landmark on the 2D image plane, as can be seen in figure 5.4, covering various facial features such as eyes, eyebrows, nose, mouth, and the contour of the face.

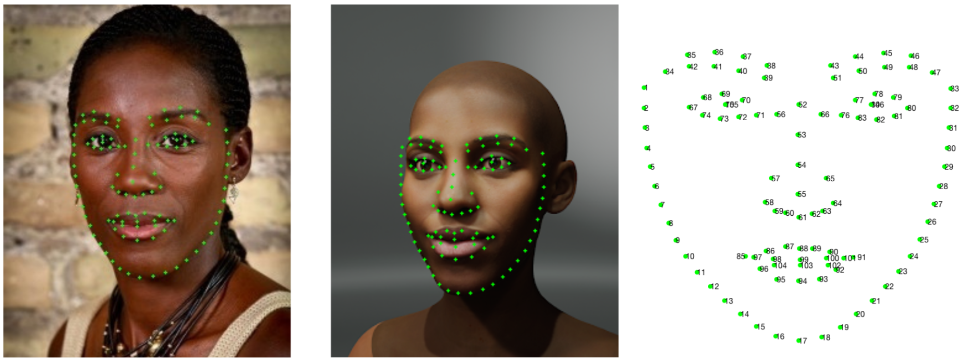


Figure 5.4: Facial landmarks delineated by InsightFace using ArcFace module, showcasing 106 key points for detailed facial feature analysis.

The following code snippet illustrates the process of employing it for extracting high-dimensional embeddings from facial images and calculating the cosine similarity between them:

```
# Path to the input images
sint1 = "<path>/1.png"
sint2 = "<path>/2.png"

# Read the first image and detect face embeddings and 2D landmarks
img1 = cv2.imread(sint1)
faces1 = app.get(img1)
face1 = faces1[0]
emb_sint1 = face1.embedding
landmark_2d_1 = face1.landmark_2d_106

# Draw 2D landmarks on the first image
for point in landmark_2d_1:
    x, y = point
    cv2.circle(img1, (int(x), int(y)), 1, (0, 255, 0), -1)

# Read the second image and detect face embeddings and 2D landmarks
img2 = cv2.imread(sint2)
faces2 = app.get(img2)
face2 = faces2[0]
emb_sint2 = face2.embedding
landmark_2d_2 = face2.landmark_2d_106

# Draw 2D landmarks on the second image
for point in landmark_2d_2:
    x, y = point
    cv2.circle(img2, (int(x), int(y)), 1, (0, 255, 0), -1)

# Calculate the CS between the embeddings of the two faces
cosine_similarity = np.dot(emb_sint1, emb_sint2) / (norm(emb_sint1)
    * norm(emb_sint2))
```

The cosine similarity metric, computed using the dot product of the embeddings divided by the product of their norms, serves as an indicator of the similarity between the two faces in the embedding space. A higher cosine similarity score suggests a closer match between the facial features represented by the embeddings.

Euclidean Distance

The **Euclidean distance** represents a measure of distance between two points by the length of the vector connecting them. In the specific context of facial recognition between images, Euclidean distance emerges as a crucial tool for assessing the similarity between vector representations of faces extracted from visual data.

Before applying Euclidean distance, facial recognition models adopt advanced feature extraction processes. These processes often involve deep neural networks for identifying and representing the salient features of a face. The result of this process is a multidimensional numerical vector that uniquely captures the peculiarities of the given face.

The Euclidean distance between two such feature vectors is calculated according to the standard formula of Euclidean geometry.

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The sum of the squared differences between corresponding components of the two vectors is then square-rooted to obtain the Euclidean distance. This process allows for a numerical evaluation of the dissimilarity or similarity between the faces represented by the vectors.

One of the main challenges of Euclidean distance in the context of facial recognition lies in its sensitivity to variations. Factors such as changes in lighting conditions, different face angles, and variations in facial expressions can significantly influence the results of Euclidean distance.
In more complex situations, relying solely on Euclidean distance may not be sufficient. Therefore, more advanced similarity metrics, such as the cosine metric, are often integrated to overcome the limitations of Euclidean distance. The robustness of such advanced metrics to environmental variations and subject conditions makes them preferable in more challenging facial recognition scenarios.

As shown in Figure 5.5, each row showcases a comparison between an original photograph of a subject (on the left) and a synthetic image generated by a Stable Diffusion model (on the right). For each pair, the Cosine Similarity and Euclidean distance was computed to quantify resemblance, also considering different poses of the subject.
Notably, the results from the Euclidean distance metric predominantly return false, except for the first image of Maria Callas. This outcome indicates that the computed distance for most pairs exceeds the threshold of 1.13, and if the Euclidean distance is greater than the threshold, the algorithm determines that the images are of different person (false). Conversely, a distance below the threshold would suggest that the images are of the same person (true). This threshold is empirically determined through machine learning techniques, specifically decision trees [68].

Contrasting these Euclidean distance findings, both the cosine similarity measures and survey responses indicate a higher perceived resemblance for the SD generated images.
In the survey, 46 participants were asked to rate the similarity of AI-generated images to the original subjects on a scale from 1 to 5. The survey results were as follows:

- For Valerio Lundini, 67.4% of participants rated the similarity as '5' and 26.1% rated it as '4'.
- For Monica Bellucci, 45.7% of participants rated the similarity as '5' and 54.3% rated it as '4'.
- For Marcello Mastroianni, 70.5% of participants rated the similarity as '5' and 29.5% rated it as '4'.
- For Maria Callas, 63.6% of participants rated the similarity as '5' and 36.4% rated it as '4'.

These ratings highlight a significant discrepancy between the quantitative Euclidean distance measures and the qualitative assessments of human observers. It was concluded that **the cosine similarity metric**, through the **ArcFace model**, was **more aligned with human perception** than Euclidean distance.

Therefore, the subsequent steps of analysis and presentation, will focus exclusively on the results obtained from this metric.

5.2 Results of Workflow Automation

Before proceeding with the analysis and comparison of Synthetic Humans derived from the selected subjects in our dataset, we will thoroughly examine some intermediate stages within the Synthetic Human creation process. This in-depth exploration will allow us to demonstrate step by step the robustness of the methods employed, culminating in the realization of the MetaHuman.

To objectively assess each of these stages, we will use the **objective metric** of **cosine similarity**, employing the facial recognition algorithm **ArcFace**.
As previously mentioned, ArcFace stands as a solid and human-perception-conforming approach to evaluate the fidelity of the generated Synthetic Humans.

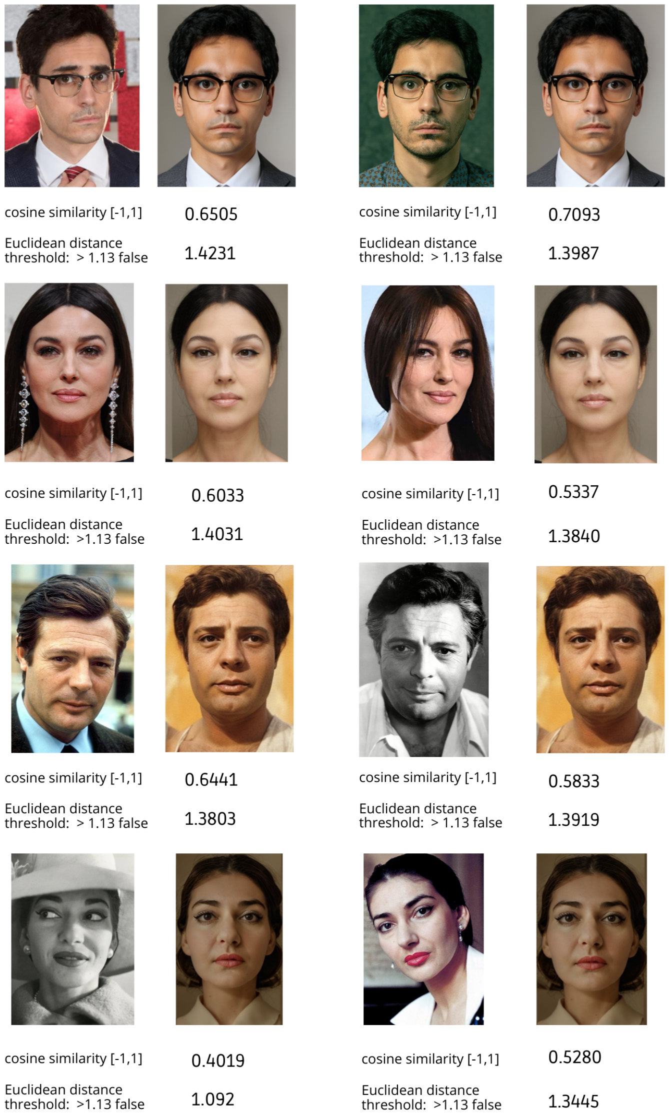

cosine similarity [-1,1]
0.6505
Euclidean distance threshold: > 1.13 false
1.4231
cosine similarity [-1,1]
0.7093
Euclidean distance threshold: > 1.13 false
1.3987
cosine similarity [-1,1]
0.6033
Euclidean distance threshold: >1.13 false
1.4031
cosine similarity [-1,1]
0.5337
Euclidean distance threshold: >1.13 false
1.3840
cosine similarity [-1,1]
0.6441
Euclidean distance threshold: > 1.13 false
1.3803
cosine similarity [-1,1]
0.5833
Euclidean distance threshold: > 1.13 false
1.3919
cosine similarity [-1,1]
0.4019
Euclidean distance threshold: >1.13 false
1.092
cosine similarity [-1,1]
0.5280
Euclidean distance threshold: >1.13 false
1.3445

Figure 5.5: Some Cosine Similarity and Euclidean distance results.

Simultaneously, we will also consider a **subjective metric** based on the results of a **survey** conducted with a sample of N participants. As mentioned before, the survey questions have been structured to reflect the same aspects examined through the objective metrics, offering a complementary perspective to the evaluation of the fidelity of the generated Synthetic Humans.

5.2.1 Validity of Images Generated through Artificial Intelligence

First and foremost, let's focus on the validation and comparison of images generated through Stable Diffusion generative network with the corresponding original images of the individual. This comparison was made using the cosine similarity metric and collecting opinions through a survey on the resemblance of the original subject to the images generated by artificial intelligence, you can observe some results in the figure 5.6.

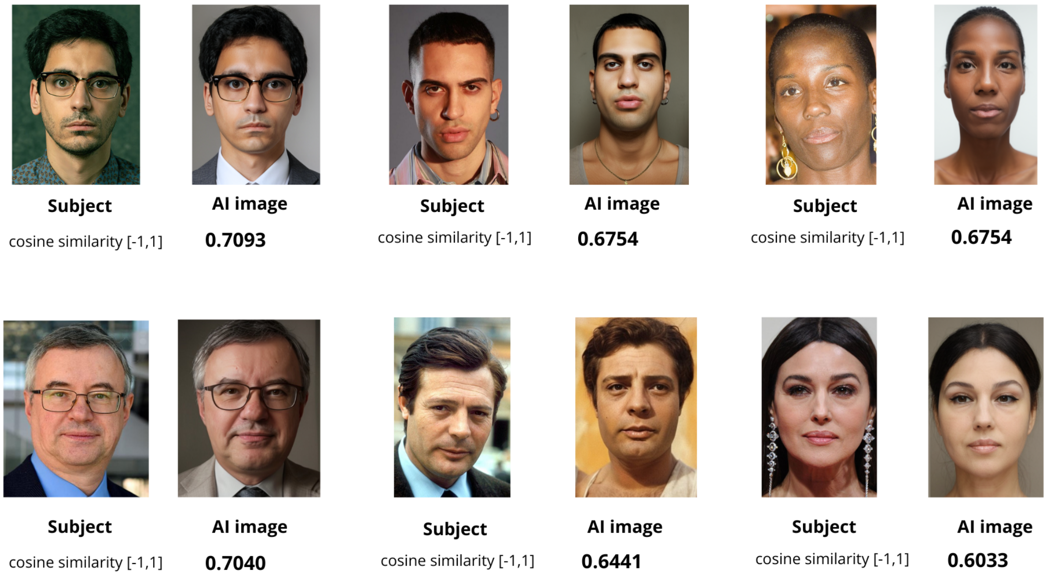


Figure 5.6: Some Results of the cosine similarity metric using ArcFace

The results obtained through the cosine similarity metric consistently exceeded 0.5, a significant threshold beyond which the probability that two subjects are the same person is very high since it suggests a high likelihood of facial match [**arcface**].

In parallel with the cosine similarity assessment, as said before we conducted a survey asking participants to *"Assign a score from 1 to 5 to express how much you think the AI-generated image resembles the original character"*. Based on the responses of 46 participants, the score of 5, which represents the highest level of resemblance, was given by approximately 68% of respondents when averaged across all subjects.

This demonstrates the remarkable capability of the generative AI in producing images that closely match the original characters in appearance.

Subject	**Cosine Similarity [-1,1]**
Alessandro Barbero	0.7040
Monica Bellucci	0.6033
Maria Callas	0.5280
Fiona May	0.6754
Marcello Mastroianni	0.6441
Mahmood	0.6754
Sadie Sink	0.6619
Valerio Lundini	0.7093
Zerocalcare	0.7065

Table 5.2: Cosine Similarity Scores of the AI generated Images

5.2.2 Validity of 3D Meshes Obtained during the Face Reconstruction Phase

In this section, we will carefully examine the validity of the 3D meshes obtained during the automated face reconstruction phase, using the script *"FB_Head_Reconstruction_Automatized.py"*.

To conduct this evaluation, we will focus on the comparison between an original image of the subject and the corresponding generated 3D mesh, maintaining the same pose as the subject in the photograph.

This comparison, similar to the previous case, was performed using the cosine similarity metric, measuring the affinity between the rendered 2D image of the 3D head mesh and the original images of the subject, you can observe the results in the figure 5.7 and in the table 5.3.
As before, we collected opinions through a survey asking *"Assign a score from 1 to 5 to express how faithful you think the 3D model of the face is to the original character"*. This was done to evaluate the perceived accuracy of the 3D facial models generated during the face reconstruction process compared to the original subjects.

The cosine similarity scores between the original 2D images and their corresponding 3D mesh reconstructions generally approached the 0.5 mark, which is indicative of a reasonable resemblance.
It's noteworthy that the slight decrease in cosine similarity when comparing 2D images to 3D meshes can be attributed to several factors like the inherent difference between a 2D image and a 3D representation. Moreover, the original images are

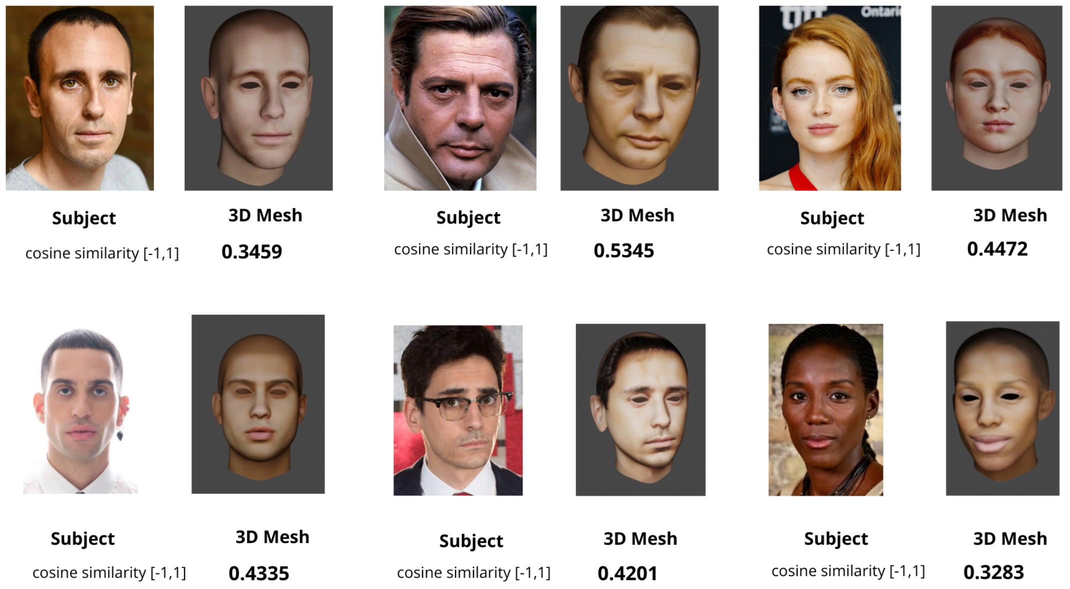


Figure 5.7: Some Results of the Cosine Similarity Metric Using ArcFace for 3D Meshes

Subject	**Cosine Similarity [-1,1]**
Alessandro Barbero	0.3545
Monica Bellucci	0.2857
Maria Callas	0.1211
Fiona May	0.3283
Marcello Mastroianni	0.5345
Mahmood	0.4335
Sadie Sink	0.4472
Valerio Lundini	0.4201
Zerocalcare	0.3459

Table 5.3: Cosine Similarity Scores of the 3D Mesh

not always captured in a frontal pose, Arcface model is susceptible to head rotation, which can impact the direct comparison. Despite these challenges, the 3-D meshes maintain a high level of fidelity to the original subjects, as the survey results also show.

In fact, the survey results give correlated results with those obtained with cosine similarity: 33.82% of participants rated the similarity with a score of 5, while a notable 29.95% assigned a score of 4. This demonstrates that a significant majority of respondents recognize a strong likeness between the 3D models and the original images

5.2.3 Validity of Synthetic Humans

In this section, we will proceed with the evaluation of the Synthetic Humans in the dataset, comparing them with an original image of the represented subject. In particular, we will test the different textures obtained from the previously analyzed processes on the MetaHuman.
This analysis aims to identify the texture creation process that produces the most satisfactory results in determining which of these procedures proves most effective in maintaining similarity to the reference subject in the MetaHuman.

This analysis, as in previous cases, was conducted using the cosine similarity metric, measuring the affinity between the face of the original subject and the face of the MetaHuman in relation to different textures. Simultaneously, we collected opinions through a survey aimed at evaluating the texture that performs better, is more realistic, and is more similar to the original subject.

Before proceeding with the analysis of the results related to the Synthetic Humans, generated by the previously examined 3D head models that have shown good performance, it is important to note that, once imported into MetaHuman, they undergo changes that can influence the face's geometry. These changes may result in the loss of specific details.

Variations in the shape of the MetaHuman's face compared to the original 3D meshes can be attributed to the "Mesh to Metahuman" process that transforms an FBX 3D mesh of a face into a MetaHuman. This process has some limitations in the customization options offered by MetaHuman Creator, which may affect the complexity and variety of replicable facial shapes. The default nature of MetaHuman Identity implies that customization options for face shape may be more limited than the richness of details contained in the original mesh. During conversion, the software tries to adapt the existing geometry to the structure of a MetaHuman, based on the available customization options, causing discrepancies in face shape during the conversion process.

After this clarification, which will partly determine the results obtained from objective metrics, let's analyze and comment on the values obtained.

Subject	Or. MH	Or. FB	Or. Blender Mix	Or. Photoshop Mix
Alessandro Barbero	0.1038	0.2688	0.1847	0.0852
Monica Bellucci	0.0832	0.3444	0.2929	0.1939
Maria Callas	0.0702	0.1760	0.1145	0.0950
Fiona May	0.0311	0.3166	0.1874	0.0651
Marcello Mastroianni	0.0687	0.3442	0.2447	0.1005
Mahmood	0.0517	0.4258	0.3413	0.1271
Sadie Sink	0.1194	0.2930	0.1879	0.0983
Valerio Lundini	0.1556	0.3447	0.3306	0.1377
Zerocalcare	0.0359	0.3108	0.2166	0.0201
Mean	0.0800	**0.3138**	0.2334	0.1025

Table 5.4: Cosine Similarity Values for Subjects Across Different Methods using Original FB and Original MH as base textures (for textures refer to the table 5.1)

Subject	Original AI	Procreate AI	Blender AI	Photoshop AI
Alessandro Barbero	0.1539	0.2084	0.1292	0.0845
Monica Bellucci	0.3155	0.3333	0.2857	0.1989
Maria Callas	0.1125	0.0957	0.0649	0.1285
Fiona May	0.2565	0.2110	0.1103	0.0912
Marcello Mastroianni	0.5506	0.4180	0.3820	0.2940
Mahmood	0.4032	0.3781	0.3150	0.1524
Sadie Sink	0.2384	0.3095	0.1737	0.1249
Valerio Lundini	0.3127	0.3480	0.2377	0.1391
Zerocalcare	0.2469	0.1945	0.1453	0.0297
Mean	**0.2878**	**0.2774**	0.2049	0.1381

Table 5.5: Cosine Similarity Values for Subjects Across Different Methods using Original AI and Original MH as base textures (for textures refer to the table 5.1)

As highlighted in Tables 5.4 and 5.5, the results obtained through the cosine similarity metric are not always good, one of the reasons for this is certainly the loss of information about face geometry once the MetaHuman is generated, as anticipated earlier.

The ***Original__FB*** 5.4 textures, stands out with an average cosine value of 0.31, which keeps most of information from the original web photos used to build the model, but results visually not particularly realistic as it is a collage of multiple photos;

The ***Procreate_AI*** and ***Original_AI*** 5.5 also show noteworthy performance, with average cosine values of 0.27 and 0.28, respectively. The latter contains all the information obtained from AI-generated images mapping, meanwhile the other one is similar but with some inaccuracies that are fixed manually on Procreate.

In the survey section, participants were presented with the task *"Select one or more images that you think represent a realistic facial texture from the available options"*, where they could choose up to three textures per subject, the render used for each subject in the survey are like the ones showed in 5.8.

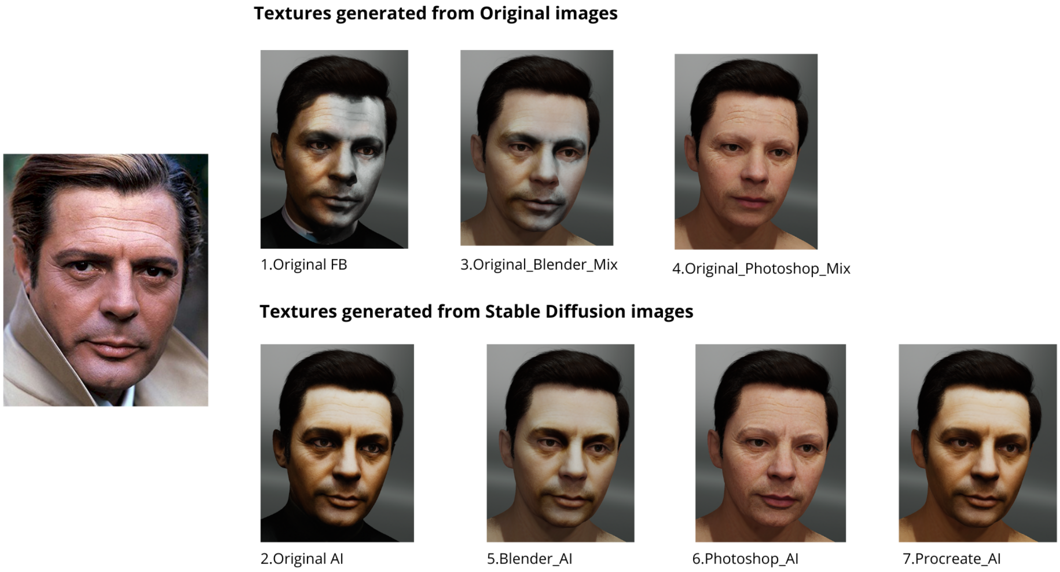


Figure 5.8: An Example of the proposed MH with different textures applied in the survey.

In the survey each MH render is presented following a random order and numbered from 1 to 7, with no reference to the specific texture used. This approach was designed to avoid biasing participants' choices, allowing an assessment solely based on visual appeal and perceived realism, rather than preconceived notions about the textures themselves.
As for the survey results, statistically, the preferred textures in terms of realism and resemblance to the original individual are Texture 6 **(Photoshop_AI)** and Texture 7 **(Procreate_AI)**.
Given that the total number of votes for this section was 690 with, Photoshop_AI

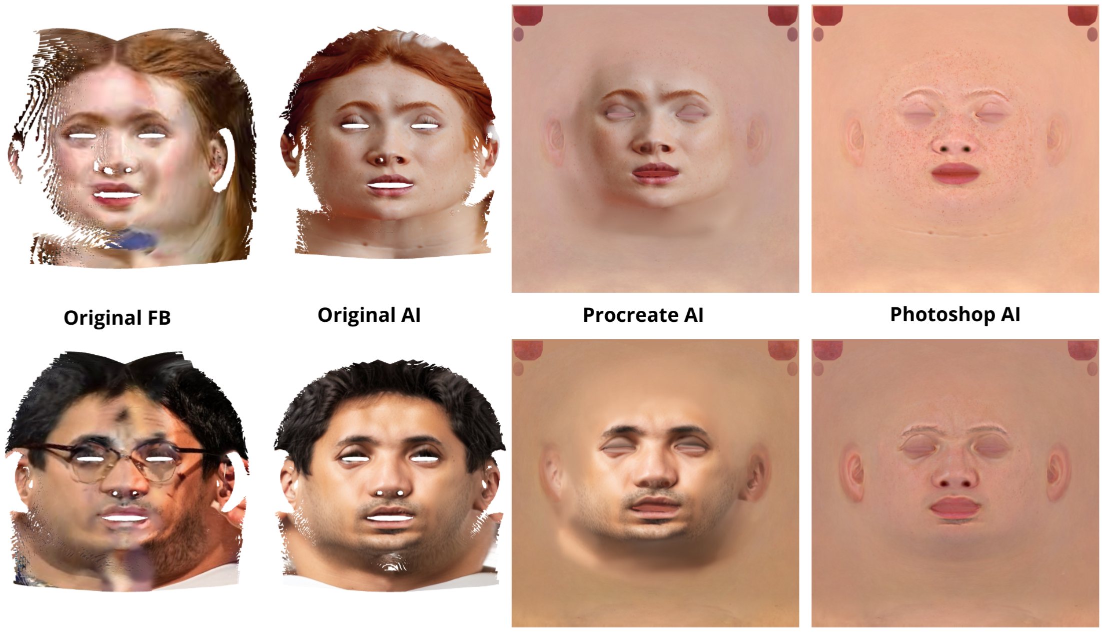


Figure 5.9: Some of the best performing facial texture maps

received 183 votes, resulting in 26.52% and Texture 7 Procreate_AI received 206 votes, leading to 29.86%.

This indicates that despite the initial higher cosine similarity of Original_FB textures in representing details from the original photos, participants showed a stronger preference for textures that realistically blend AI-generated details with manual corrections.

Also, it emerges that the texture that ranks among the best in both metrics is **Procreate_AI.**

Metahuman Creator VS Character Creator 3: Comparative Analysis

In the context of creating photorealistic avatars, **MetaHuman Creator** and **Character Creator 3** [36] stand out as the main options.

In light of the not entirely satisfactory performance of MetaHuman, Character Creator 3 was considered since it's renowned for its features and flexibility, presented itself as a promising alternative.
The avatars resulting from the creation process with Character Creator 3 (CC3) underwent the same evaluation protocol as used earlier, ensuring a consistent and fair assessment across different avatar generation platforms.
In contrast to MH, which are shown to the voter "wearing" the Procreate_AI facial texture, CC3 employs an automated texture generation approach. This process utilizes the original AI-generated photos of the subjects, produced by Stable Diffusion, to create the textures. Furthermore, CC3 avatars have a facial UV mapping that differs from the MH one, as shown in Figure 5.10. This automatic texture generation in CC3, starting directly from the subject's AI-generated images, contrasts sharply with the manual texture fixing approach used in MH.

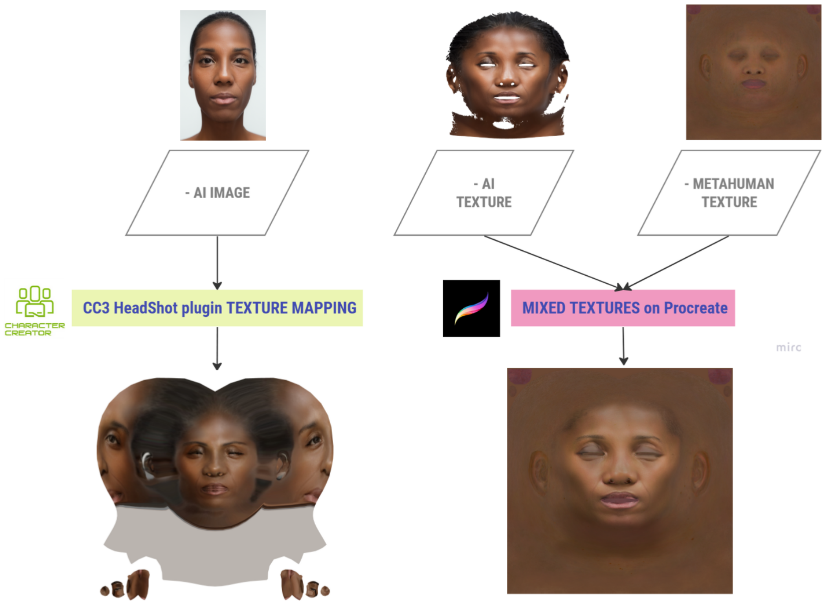

- AI IMAGE
- AI TEXTURE
- METAHUMAN TEXTURE
CHARACTER CREATOR
CC3 HeadShot plugin TEXTURE MAPPING
MIXED TEXTURES on Procreate
mirc

Figure 5.10: Displayed side-by-side, the distinct approaches to facial texture mapping: CC3's methodology is showcased on the left, while MH's technique is on the right

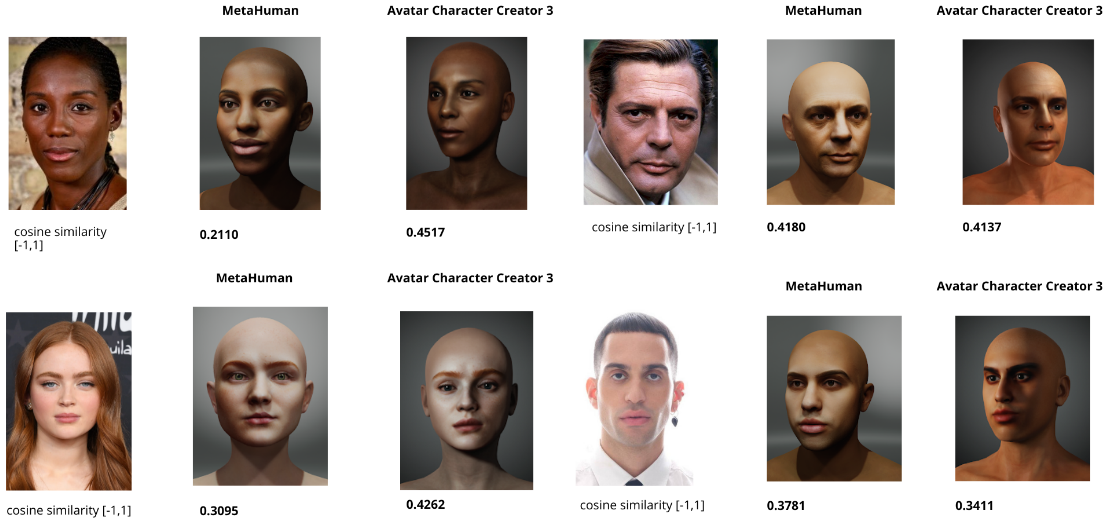


Figure 5.11: Some Results of the Cosine Similarity Metric Using ArcFace for MetaHumans and Avatars Obtained with Character Creator 3.

The results derived from the cosine similarity metric, shown in Table 5.6 and in figure 5.11, reveal that avatars generated using Character Creator often demonstrate comparable or higher values in comparison to those produced by MetaHuman.

Subject	**MetaHuman**	**Character Creator 3 Avatar**
Alessandro Barbero	0.2083	0.2394
Monica Bellucci	0.3136	0.3817
Maria Callas	0.0957	0.2210
Fiona May	0.0211	0.4517
Marcello Mastroianni	0.4180	0.4137
Mahmood	0.3781	0.3411
Sadie Sink	0.3095	0.4262
Valerio Lundini	0.3480	0.4417
Zerocalcare	0.1945	0.5108
Mean	**0.2541**	**0.3808**

Table 5.6: Cosine Similarity Scores of the Final Synthetic Humans Mesh

Nevertheless, the conducted survey asking *"Which of the two Avatars most looks like the original subject?"* revealed mixed preferences among participants regarding Synthetic Humans presented as MetaHuman (MH) and Character Creator (CC3) avatars.
Out of 414 total responses (46 voters each choosing between two avatar types for 9 subjects), avatars created with Character Creator were preferred in approximately 57.73% (239 votes) of instances, while MetaHuman Creator was favored in 42.27% (175 votes). This preference distribution is indicating a inclination towards CC3's avatars, yet also revealing a significant appreciation for MH's ones.

Despite the variability in user preferences, this information has enabled us to conduct a thorough comparison between the performance of Character Creator 3 and MetaHuman, marking a significant stride in the quest for more suitable solutions. This lays the groundwork for further investigation and enhancements in the Synthetic Human creation process, outlining avenues for the future evolution of this technology.

5.2.4 Results Considerations

The detailed analysis conducted in the Model Validation chapter has brought to light several key aspects regarding the effectiveness of our automated workflow in generating Synthetic Humans. Before examining the results in detail, it is essential to emphasize the importance of our generative network-based approach.

The use of generative networks for subject image generation has proven to play a significant role in improving the quality and similarity of Synthetic Humans compared to traditional methods. Generative networks allow capturing complex details and nuances present in the original images, producing realistic textures that maintain fidelity to the reference individual.

Briefly summarizing the main results obtained from different phases of the process:

1. **Validity of AI-Generated Images:**
   Images generated by generative networks have proven to be valid and realistic, with cosine similarity values consistently above 0.5, confirming their resemblance to the original counterparts. Survey results further supported the high fidelity of the generated images.

2. **Validity of 3D Meshes Obtained during Face Reconstruction:**
   3D meshes obtained during the face reconstruction phase using the script *"FB_Head_Reconstruction_Automatized.py"* maintained a high level of fidelity to the reference individual. Cosine similarity confirmed consistency between 3D meshes and original images.

3. **Validity of Synthetic Humans:**
   Cosine similarity metric results for Synthetic Humans showed a variety of values, with a tendency to have higher values in textures preserving more details of the original face. The survey highlighted a subjective preference for textures obtained through *Procreate_AI* and *Photoshop _AI*.

4. **Comparing MetaHuman Creator and Character Creator 3:**
   Character Creator 3 has proven to be a promising alternative to MetaHuman Creator, with cosine similarity results often similar or superior. However, the survey did not reveal a clear user preference for either approach.

The generative network-based approach, therefore, offers greater realism in textures, preserving specific details of the original faces. However, it is important to note that some challenges persist, especially during the 3D mesh-to-MetaHuman conversion phase within the Metahuman Creator workflow, potentially resulting in the loss of geometric facial information.

Although MetaHumans produced via semi-automated workflows are rapidly created, they have not yet attained the nuanced quality of 3D models of Synthetic Humans meticulously handcrafted or partially manually sculpted.
Nevertheless, this methodology establishes a reliable starting point for the development of 3D meshes and facial texturing.
While our automated methods provide a substantial starting point, manual refinement, through targeted sculpting of the 3D model and meticulous texture adjustments, can elevate the final product to an even higher standard of realism. The approach proposed in this thesis serves as a substitute for the more tedious manual tasks, streamlining the process significantly, especially valuable in crafting the subject's human head, a critical element for recognition.
These results provide a solid foundation for further research and developments in Synthetic Human creation.

Chapter 6

Conclusions

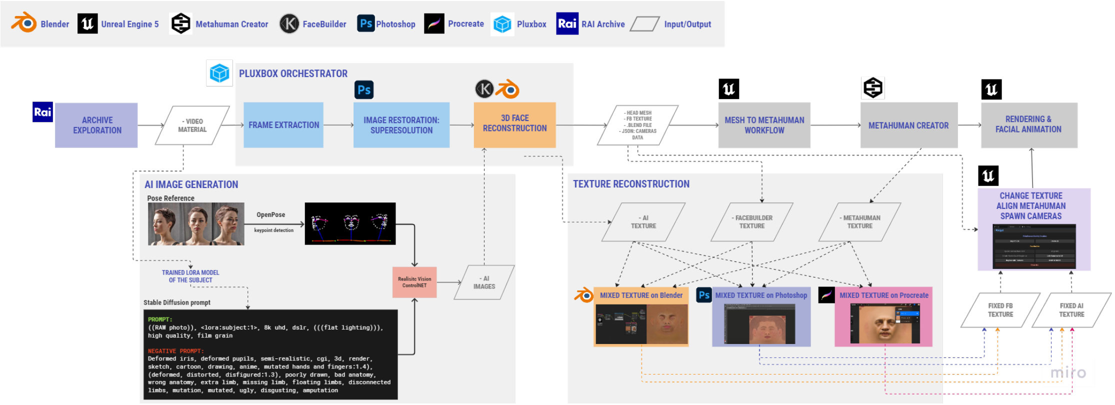
6.1 Summary and Final Considerations

This thesis, conducted in close collaboration with **RAI R&D**, represents a significant step towards the optimization and automation of the photorealistic avatar creation process in the context of the broadcast industry. The main objective was to explore and implement an advanced workflow for the generation of Synthetic Humans, aiming to minimize human intervention and ensure an efficient and flexible process.

The motivation behind this thesis was the challenge of revitalizing RAI's extensive archive of images and videos, transforming it into a valuable resource for the creation of photorealistic 3D avatars. The evolution of the broadcast industry demands increasingly engaging and interactive content, and the transformation of two-dimensional resources into photorealistic avatars of significant personalities, such as historical figures or past celebrities, offers an innovative approach to enhancing television programming with virtual duets, interviews, shows, and more.

As shown in Figure 6.1, illustrating the implemented workflow in this thesis, the initial phase focused primarily on the complete automation of the Face Reconstruction stage. This was made possible by the Orchestrator, developed in collaboration with PluxBox, a partner of RAI in the IBC2023 Accellerator Project.

The Orchestrator, a sophisticated technological solution, played a central role in integrating various stages into a single block, from image selection to super resolution, and finally, three-dimensional face reconstruction. The automation of Face Reconstruction was facilitated by a script, made available to the Orchestrator through Rest APIs, utilizing the functionalities of the Blender FaceBuilder plugin to generate a 3D face mesh from 2D images.



```
PROMPT:
((RAW photo)), <lora:subject:1>, 8k uhd, dslr, (((flat lighting))),
high quality, film grain

NEGATIVE PROMPT:
Deformed iris, deformed pupils, semi-realistic, cgi, 3d, render,
sketch, cartoon, drawing, anime, mutated hands and fingers:1.4),
(deformed, distorted, disfigured:1.3), poorly drawn, bad anatomy,
wrong anatomy, extra limb, missing limb, floating limbs, disconnected
limbs, mutation, mutated, ugly, disgusting, amputation
```

Figure 6.1: Implemented Workflow for Synthetic Human Creation

Subsequently, emphasis was placed on improving textures through the adoption of Stable Diffusion to generate images of the reference subject with the help of Artificial Intelligence. The use of AI-generated images proved more suitable for texture production than traditional methods. This phase was influenced by the refinement of LoRA training models, enabling the generation of realistic and detailed images of the target subjects, crucial for creating uniform textures in the Synthetic Human generation process.

The in-depth analysis conducted in the Model Validation chapter aims to assess the effectiveness of various steps within the workflow.
Images created through generative neural networks demonstrated significant similarity to their original counterparts, supported not only by cosine similarity metrics but also by survey results confirming the high fidelity of the generated images.
The fidelity of 3D meshes obtained during the automated Face Reconstruction phase remained at a high level, confirmed by both objective metrics and survey values.
However, the similarity comparison phase between the obtained MetaHuman and the original reference subject presented some critical issues.

Considering the previous data, it is evident that the use of generative neural networks for rendering subject images allows for greater realism in textures, preserving specific details of the original faces. Moreover, the validation of the automation part of 3D mesh creation from images represents a significant step forward in the automation process of Synthetic Human creation. However, the studied process is not without challenges, as it was not possible to fully automate the entire workflow due to limitations in Metahuman APIs, even though the implementation of Widgets within Unreal Engine allowed for semi-automation of this phase. The main issue encountered relates to the conversion of 3D mesh to MetaHuman, where some cases experience a loss of facial geometry information.

Nevertheless, these results provide a starting point for further research, exploration, and developments in Synthetic Human creation.

6.2 Future Developments

Looking to the future, this thesis lays the groundwork for significant developments in the broadcast industry, marking a substantial shift in the approach to photorealistic avatar creation. The automation of Synthetic Human generation is just the beginning of a potentially revolutionary transformation in audience interaction with visual content, paving the way for extensive integration into broadcasting, the metaverse, and video games contexts. However, to reach this stage, it is imperative to dedicate further efforts to in-depth studies, testing, and the exploration of new technologies, emphasizing the progressive nature and the early development path of this ambitious goal.

The following are potential directions for future developments, for further improvements in the Synthetic Human generation process and related automations. This analysis aims to identify key directions for subsequent research and development, outlining opportunities that can contribute to further consolidating and refining photorealistic avatar creation in the broadcast industry.

6.2.1 Optimization of the MetaHuman Conversion Process

A potential future development involves optimizing the conversion process from 3D mesh to MetaHuman. The current process, following the MetaHuman plugin workflow for Unreal Engine, known as ***Mesh To MetaHuman***, offers an intuitive solution, but variations in facial geometry may occur during the conversion to Metahuman. This variation is attributed to limitations in customization options offered by MetaHuman Creator during the conversion process.

Due to this issue, after executing the "Mesh to Metahuman" process, it is recommended to explore advanced customization options provided by MetaHuman Creator. This post-conversion step offers the possibility to further refine facial shape through sculpting tools and manual adjustment features within Metahuman Creator. Navigating through these advanced options can be crucial for adapting MetaHuman to specific aesthetic requirements, allowing for more detailed facial customization.

However, it is essential to note that if MetaHuman Creator has a limited range of customization options for facial shape compared to your original mesh, discrepancies in facial shape may still occur. In other words, MetaHuman Creator may not precisely replicate the face shape of your original mesh. The default nature of MetaHuman Identity implies that the variety of customizable facial shapes may be limited compared to the complexity of the original mesh.

Use of External 3D Sculpting Tools

In cases where MetaHuman Creator does not fully meet desired customization needs, an alternative approach is the use of external 3D sculpting tools. Through this methodology, the MetaHuman mesh can be exported and subsequently imported into platforms such as ZBrush, Blender, or Maya. These software, known for their advanced facial geometry manipulation capabilities, can offer an additional level of control over facial shape, allowing for refined modeling and more detailed adaptation to the original mesh concept.

A particularly relevant tool for Metahuman customization in Maya is **Metapipe (Custom Metahuman & Expressions Tool)** [69] [70].
Metapipe, integrated directly into Maya, provides 3D artists with a powerful tool for advanced Metahuman customization, positioning itself as a promising starting point for future developments aimed at overcoming limitations in the "Mesh to Metahuman" workflow of MetaHuman Creator, offering a level of control and flexibility beyond current constraints. It allows complete control over the mesh, an automated workflow, and the use of official Epic Games Calibration DNA Codes, significantly simplifying the customization process. Therefore, Metapipe appears as a potential future exploration, as it features characteristics that enable overcoming constraints in the standard "Mesh to Metahuman" process.

6.2.2 Exploration of Alternative Solutions to MetaHuman

Considering the challenges and limitations identified in the Metahuman workflow, such as the loss of geometric information during conversion, along with automation process restrictions due to the lack of APIs for remote use of Metahuman services, the exploration of alternative software for Photorealistic Avatar creation, such as **Character Creator 4** [36], could be evaluated.
In the latest update, Character Creator 4, augmented by a newly released Headshot plugin version, introduces a workflow similar to the 'Mesh to MetaHuman' process. Unlike the previous version, it directly uses the imported 3D mesh rather than solely depending on 2D image data. As demonstrated in Figure 2.12, Character Creator offers smooth integration with leading industry tools, enabling users to import characters seamlessly into their projects. Although Character Creator does

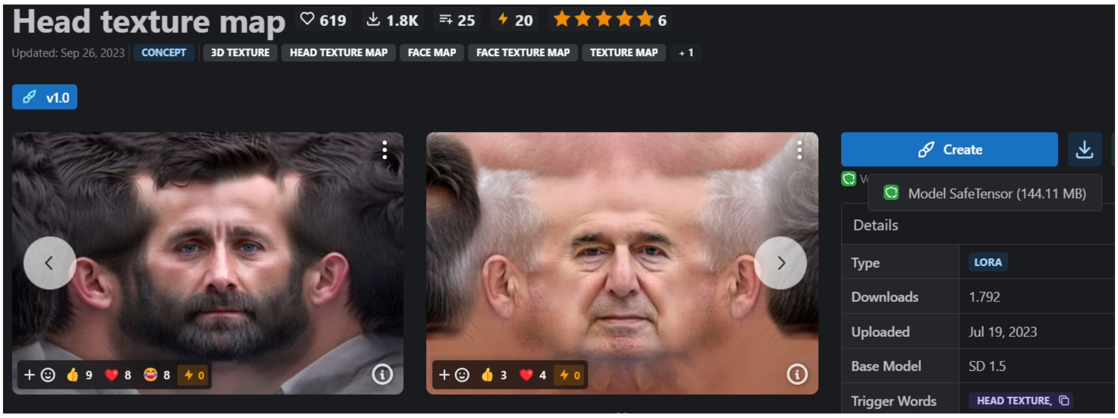
not currently provide tools to complete automation, it could represent a fertile area for further developments, offering an alternative to the existing process.

This option is also motivated by the results in the Model Validation chapter, where both the survey and cosine similarity metric indicated values often superior between the original subject and the avatar generated by Character Creator 3 compared to those obtained with avatars generated by Metahuman. Exploring this direction could lead to significant improvements in Synthetic Human creation, reducing limitations observed in the current process.

6.2.3 Optimizations in the Use of Generative Networks

Optimizing generative networks could allow even more accurate rendering of textures, eliminating the need for manual post-generation modifications. Exploring new training models or advanced generation techniques could contribute to achieving increasingly convincing and faithful results to the originals.

In the pursuit of enhancing generative networks for faster texture rendering, in Figure 6.2 the **"Head Texture Map"** [71] LoRA model, presented on Civitai, offers a notable example. Based on Stable Diffusion 1.5, this model has been fine-tuned for the specific task of generating facial textures. Its size is only 144MB and uses trigger words such as *Head texture* and *<lora:Head Texture Map_v.1.0:1>*. The produced Facial UV map, is not yet compatible with MetaHuman's facial UV mapping, but it signifies a step towards more automated, AI-driven texture that may soon be tailored to project-specific requirements, potentially bypassing traditional tools like FaceBuilder for texture creation.



Head texture map

Updated: Sep 26, 2023

Model SafeTensor (144.11 MB)

Details

Type	LORA
Downloads	1.792
Uploaded	Jul 19, 2023
Base Model	SD 1.5
Trigger Words	HEAD TEXTURE,

Figure 6.2: The Civitai web interface showcasing the 'Head Texture Map' model available for download

Another future development is the advent of **PanoHead** [72], a pioneering 3D generative model that synthesizes and reconstructs comprehensive 3D human heads from any viewpoint. PanoHead's innovative approach in using unstructured 2D images for training, while maintaining 3D consistency across broad viewing angles, signifies a substantial evolution beyond the capabilities of traditional 3D GANs. This could revolutionize digital human modeling by creating detailed and diverse 3D heads, including intricate hairstyles, potentially bypassing the need for manual texturing and 3D sculpitng processes.

Integrating AI Texture Generation with the Orchestrator

The potential to integrate AI-driven texture generation into scripted workflows is a compelling research and development frontier. Pluxbox's orchestrator, renowned for its powerful integrations and open API architecture, could potentially facilitate seamless interactions with a local deployment of Stable Diffusion's API. The synergy between these platforms could significantly streamline the Synthetic Human production process, not only by enhancing the 3D head reconstruction process from raw video data but also integrating consistent facial texture. Another scenario could be to integrate AI texture generation directly inside Blender after the FaceBuilder phase. Stability AI provides a dedicated integration with Blender [73]. This allows users to directly apply AI-driven image generation and manipulation within Blender, offering capabilities such as Image-to-Image transformations on rendered frames or the creation of textures from textual descriptions.

Despite the analysis of these potential future developments, as we are still in the early stages of this project, aiming to create a fully automated and highly efficient workflow for broadcast industry adoption, it is essential to adopt a continuous evaluation and adaptation approach. This approach involves constant monitoring of technological developments and flexibility in modifying the workflow in response to new opportunities and challenges in the technological and industrial landscape.

In conclusion, this study not only establishes a base for future massive integration of Synthetic Humans in media and virtual contexts but also opens new perspectives for visual storytelling and interactive content experiences. Continuous commitment to research and development will fully leverage the potential of this technology, contributing to shaping the future of the broadcast industry.

Bibliography

[1] *Human Digital Twins: Creating New Value Beyond the Constraints of the Real World.* https://www.rd.ntt/e/ai/0004.html. 2023 (cit. on p. 3).

[2] *An Era of Digital Humans: Pushing the Envelope of Photorealistic Digital Character Creation.* https://developer.nvidia.com/blog/an-era-of-digital-humans-pushing-the-envelope-of-photorealistic-digital-character-creation/. 2021 (cit. on p. 2).

[3] Gillen McAllister. *The story behind The Last of Us Part 2's staggeringly realistic in-game character facial animation.* PlayStation.Blog. 2020. URL: https://blog.playstation.com/2020/08/28/the-story-behind-the-last-of-us-part-iis-staggeringly-realistic-in-game-character-facial-animation/ (cit. on p. 3).

[4] John Linneman. *Detroit: Become Human is a different kind of tech showcase.* https://www.eurogamer.net/digitalfoundry-2018-detroit-become-human-tech-analysis. 2018 (cit. on p. 3).

[5] THE THIRD FLOOR, Inc. *Virtual Visualization Series – The Mandalorian.* https://thethirdfloorinc.com/4206/virtual-visualization-series-the-mandalorian/. Online; accessed November 17, 2023. 2020 (cit. on p. 3).

[6] lilmiquela. *Miquela's Instagram Profile.* https://www.instagram.com/lilmiquela/. Instagram profile. 2023 (cit. on p. 4).

[7] *ABBA Voyage Official Website - 2023 ABBA Concert in London.* https://abbavoyage.com/ (cit. on p. 4).

[8] Brian Caulfield. *NVIDIA, BMW Blend Reality, Virtual Worlds to Demonstrate Factory of the Future.* https://blogs.nvidia.com/blog/nvidia-bmw-factory-future/. 2021 (cit. on p. 4).

[9] Lex Fridman. *Transcript for Mark Zuckerberg: First Interview in the Metaverse.* https://lexfridman.com/mark-zuckerberg-3-transcript. 2023 (cit. on p. 6).

[10] Amelia Tait. «'Here is the news. You can't stop us': AI anchor Zae-In grants us an interview». In: *The Guardian* (Oct. 2023). URL: https://www.theguardian.com/tv-and-radio/2023/oct/20/here-is-the-news-you-cant-stop-us-ai-anchor-zae-in-grants-us-an-interview (cit. on p. 6).

[11] *Digital Human.* https://www.alibabacloud.com/solutions/digital-human. Alibaba Cloud, 2023 (cit. on p. 6).

[12] Jake Bickerton. «Digital Human created for Tour de France». In: *Broadcast Now* (June 2023). URL: https://www.broadcastnow.co.uk/production/digital-human-created-for-tour-de-france/5183644.article (cit. on p. 6).

[13] *Digital Humans.* https://triplesense.it/digital-humans. Triplesense Reply, 2023 (cit. on p. 7).

[14] Lu Chen, Sida Peng, and Xiaowei Zhou. «Towards efficient and photorealistic 3D human reconstruction: A brief survey». In: (2021). ISSN: 2468-502X. DOI: https://doi.org/10.1016/j.visinf.2021.10.003. URL: https://www.sciencedirect.com/science/article/pii/S2468502X21000413 (cit. on pp. 7–9).

[15] *Adobe Super Resolution.* https://www.adobe.com/products/photoshop-lightroom/super-resolution.html. n.d. (Cit. on p. 13).

[16] *Topaz Labs Gigapixel AI.* https://www.topazlabs.com/gigapixel-ai. n.d. (Cit. on p. 13).

[17] Zitong Wang, Jue Zhang, Ting Chen, Wen Wang, and Ping Luo. «Restore-Former++: Towards Real-World Blind Face Restoration From Undegraded Key-Value Pairs». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). https://doi.org/10.1109/tpami.2023.3315753, pp. 1–15 (cit. on p. 13).

[18] Chongyi Li Shangchen Zhou Kelvin C.K. Chan. «Towards Robust Blind Face Restoration with Codebook Lookup Transformer». In: (). https://arxiv.org/pdf/2206.11253.p (cit. on p. 14).

[19] *Intel RealSense.* https://www.intelrealsense.com/. n.d. (Cit. on p. 15).

[20] *Capturing Reality.* https://www.capturingreality.com/. n.d. (Cit. on p. 15).

[21] *Skanect by Structure.* https://structure.io/skanect. n.d. (Cit. on p. 16).

[22] «FaceBuilder for Blender | KeenTools». In: (n.d.). https://keentools.io/products/facebuilder-for-blender (cit. on p. 16).

[23] KeenTools. «FaceBuilder for Blender Guide». In: (2021). https://medium.com/keentools/facebuilder-for-blender-guide-cbb10c717f7c (cit. on pp. 16, 17).

[24] *Reallusion Character Creator - Headshot*. https://www.reallusion.com/character-creator/headshot/. n.d. (Cit. on p. 17).

[25] *Daz 3D - Face Transfer Unlimited*. https://www.daz3d.com/face-transfer-unlimited. n.d. (Cit. on p. 18).

[26] *Avatar SDK*. https://avatarsdk.com/. n.d. (Cit. on p. 18).

[27] *FaceGen*. https://facegen.com/. n.d. (Cit. on p. 18).

[28] Apostolos Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vassilis Triantafyllou, Aneesh Ghosh, and Stefanos Zafeiriou. «AvatarMe: Realistically Renderable 3D Facial Reconstruction 'in-the-wild'». In: (n.d.). https://arxiv.org/pdf/2003.13845.pdf (cit. on pp. 18, 19).

[29] Taras Martyniuk, Oleksandr Kupyn, Yurii Kurlyak, Ivan Krashenyi, Jiri Matas, and Viktoriia Sharmanska. «DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image». In: (n.d.). https://arxiv.org/pdf/2204.03688.pdf (cit. on p. 19).

[30] Baris Gecer, Jia Deng, and Stefanos Zafeiriou. «OSTeC: One-Shot Texture Completion». In: (n.d.). https://arxiv.org/pdf/2012.15370.pdf (cit. on pp. 19, 20).

[31] *Reallusion iClone*. https://www.reallusion.com/iclone/. n.d. (Cit. on p. 21).

[32] *NVIDIA Omniverse - Audio2Face*. https://www.nvidia.com/it-it/omniverse/apps/audio2face/. n.d. (Cit. on p. 21).

[33] *FacewareTech*. https://facewaretech.com/. n.d. (Cit. on p. 21).

[34] *Unreal Engine - MetaHuman*. https://metahuman.unrealengine.com/. n.d. (Cit. on p. 22).

[35] «New MetaHuman Animator Feature Set to Bring Easy High-Fidelity Performance Capture to MetaHumans». In: (n.d.). https://www.unrealengine.com/en-US/blog/new-met (cit. on p. 22).

[36] *Character Creator*. https://www.reallusion.com/character-creator/. n.d. (Cit. on pp. 23, 67, 74).

[37] *Stable Diffusion*. https://stability.ai/stable-diffusion (cit. on p. 25).

[38] NVIDIA. *Generative AI – What is it and How Does it Work?* https://www.nvidia.com/en-us/glossary/data-science/generative-ai/. NVIDIA's explanation and overview of Generative AI. 2023 (cit. on p. 25).

[39] *Generative Adversarial Network (GAN)*. https://lilianweng.github.io/posts/2017-08-20-gan/#generative-adversarial-network-gan. 2017 (cit. on p. 26).

[40] *This Person Does Not Exist*. https://thispersondoesnotexist.com/ (cit. on p. 26).

[41] Google Developers. *Common Problems in Machine Learning*. https://developers.google.com/machine-learning/gan/problems. A guide to common problems in Generative Adversarial Networks (GANs) on Google Developers platform. 2022 (cit. on p. 26).

[42] *Denoising Diffusion Probabilistic Models*. https://ar5iv.org/pdf/2006.11239.pdf. 2020 (cit. on p. 26).

[43] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. «Diffusion Models in Vision: A Survey». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14.8 (Aug. 2022). Available at arXiv: https://arxiv.org/pdf/2209.04747.pdf, p. 1 (cit. on p. 26).

[44] Google Cloud Tech. *Introduction to image generation*. https://www.youtube.com/watch?v=kzxz8CO_oG4. 2023 (cit. on p. 26).

[45] Steins Fu. *Stable Diffusion Clearly Explained*. https://medium.com/@steinsfu/stable-diffusion-clearly-explained-ed008044e07e. 2023 (cit. on pp. 27, 28).

[46] «You Can't Spell Diffusion Without U». In: *Towards Data Science* (2023). URL: https://towardsdatascience.com/you-cant-spell-diffusion-without-%20u-60635f569579 (cit. on p. 27).

[47] Robin Rombach and Björn Ommer. «High-Resolution Image Synthesis with Latent Diffusion Models». In: *arXiv preprint arXiv:2112.10752* (2021). URL: https://ar5iv.org/abs/2112.10752 (cit. on p. 28).

[48] *How Stable Diffusion Work - Latent Diffusion Model | Stable Diffusion Art*. 2023. URL: https://stable-diffusion-art.com/how-stable-diffusion-work/#Latent_diffusion_model (cit. on p. 28).

[49] *CLIP - Transformers - Hugging Face Documentation*. URL: https://huggingface.co/docs/transformers/model_doc/clip (cit. on p. 29).

[50] *DALL · E 2*. https://openai.com/dall-e-2. OpenAI, 2023 (cit. on p. 29).

[51] *Midjourney*. https://www.midjourney.com/home. 2023 (cit. on p. 29).

[52] AUTOMATIC1111. *Stable Diffusion web UI*. https://github.com/AUTOMATIC1111/stable-diffusion-webui. A browser interface based on Gradio library for Stable Diffusion. 2023 (cit. on p. 30).

[53] *AUTOMATIC1111 - Stable Diffusion Art*. 2023. URL: https://stable-diffusion-art.com/automatic1111/ (cit. on p. 31).

[54] Civitai. *Realistic Vision V5.1 - V5.1 (VAE) | Stable Diffusion Checkpoint.* `https://civitai.com/models/4201/realistic-vision-v51`. 2023 (cit. on p. 31).

[55] *LoRA: Low-Rank Adaptation of Large Language Models.* `https://arxiv.org/abs/2106.09685` (cit. on p. 32).

[56] *LoRA - Stable Diffusion Art.* URL: `https://stable-diffusion-art.com/lora/` (cit. on p. 32).

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. «Adding Conditional Control to Text-to-Image Diffusion Models». In: *arXiv:2302.05543 [cs.CV]* (2023). DOI: `10.48550/arXiv.2302.05543`. URL: `https://doi.org/10.48550/arXiv.2302.05543` (cit. on p. 33).

[58] lllyasviel. *sd-controlnet-openpose - Hugging Face Model.* Hugging Face Model Hub. URL: `https://huggingface.co/lllyasviel/sd-controlnet-openpose` (cit. on p. 33).

[59] *ControlNet - Stable Diffusion Art.* 2023. URL: `https://stable-diffusion-art.com/controlnet/` (cit. on p. 33).

[60] Epic Games. *Widget Blueprints in UMG for Unreal Engine.* `https://docs.unrealengine.com/5.2/en-US/widget-blueprints-in-umg-for-unreal-engine/`. 2023 (cit. on p. 34).

[61] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.* 2022. arXiv: `2201.12086` (cit. on p. 39).

[62] AUTOMATIC1111. *Features - Attention/Emphasis | Stable Diffusion WebUI Wiki.* `https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Features#attentionemphasis`. 2023 (cit. on p. 40).

[63] *3D Scan Store.* `https://www.3dscanstore.com/` (cit. on p. 44).

[64] *Pluxbox Platform.* `https://pluxbox.com/pluxbox-platform/` (cit. on p. 45).

[65] X. Wei, H. Wang, B. Scotney, and H. Wan. «CosFace: Large Margin Cosine Loss for Deep Face Recognition». In: *Pattern Recognition* 97 (2020), p. 107012. DOI: `10.1016/j.patcog.2019.107012` (cit. on p. 54).

[66] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. *ArcFace: Additive Angular Margin Loss for Deep Face Recognition.* cite arxiv:1801.07698Comment: ArcFace with parallel acceleration. 2022. URL: `http://arxiv.org/abs/1801.07698` (cit. on p. 54).

[67] InsightFace Developers. *InsightFace: 2D and 3D Face Analysis Project.* `https://github.com/deepinsight/insightface`. 2023 (cit. on p. 55).

[68] Sefik Ilkin Serengil. *Deep Face Recognition with ArcFace in Keras and Python.* https://sefiks.com/2020/12/14/deep-face-recognition-with-arcface-in-keras-and-python/. 2020 (cit. on p. 57).

[69] *Road to Custom MetaHuman: MetaPipe Custom MetaHuman Expressions Tool.* `https://www.artstation.com/marketplace/p/PmpWr/road-to-custom-metahuman-metapipe-custom-metahuman-expressions-tool`. n.d. (Cit. on p. 74).

[70] *Arts and Spells Documentation.* `https://www.artsandspells.com/documentation`. n.d. (Cit. on p. 74).

[71] *Head texture map - v1.0 | Stable Diffusion LoRA.* `https://civitai.com/models/112287/head-texture-map` (cit. on p. 75).

[72] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. *PanoHead: Geometry-Aware 3D Full-Head Synthesis in 360deg.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2023 (cit. on p. 75).

[73] *Stability for Blender.* `https://platform.stability.ai/docs/integrations/blender` (cit. on p. 76).